DEVELOPING BOTTOM-UP, INTEGRATED OMICS METHODOLOGIES FOR BIG DATA BIOMARKER DISCOVERY

Bobak David Kechavarzi

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics and Computing,
Indiana University

November 2020

Accepted by the Graduate Faculty of Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

_____
Huanmei Wu, PhD, Chair

_____
Thompson Doman, PhD

November 22, 2019

_____
Ernst Dow, PhD

_____
Yunlong Liu, PhD

_____
Xiaowen Liu, PhD

_____
Jingwen Yan, PhD

DEDICATION

For those who came before to pave the way, and those who will come to go farther.

ACKNOWLEDGEMENT

Practitioners and students of the life sciences describe themselves as “lifelong learners.” Many have helped me in scholastic, personal, and professional development along my journey thus far.

Drs. Hughes and Rubino, thank you for being the catalysts in my desire to pursue the sciences. From your undergraduate courses at Hanover I found my enjoyment of genetics and biostatistics which provided the bedrock of my Bioinformatics pursuits. Without these skills I would not have gotten far.

Dr. Šabanović, your dedication and peerless mentorship set the standard for research excellence. Thank you so much for the opportunity to work and grow within your group. Those experiences were the compass for my efforts in research and leadership through the times that followed.

Drs. Janga, Wu, and the faculty and committee supporting me; thank you for your time and direction as my PIs, chair, and mentors while at IUPUI. Through our work and collaboration I have developed my investigative desires and matured my independent research skills. I’d like to extend a special thank you to Dr. Doman; for acting as co-chair while also managing his responsibilities at Eli Lilly, your input was invaluable. Chapter 1 was adapted from: Kechavarzi, Bobak, Janga, Sarath. Dissecting the expression landscape of RNA-binding proteins in human cancers. Genome Biology 201415:R14.https://doi.org/10.1186/gb-2014-15-1-r14. Chapter 2 was adapted from: Kechavarzi BD, Wu H, Doman TN. Bottom-up, integrated -omics analysis identifies broadly dosage-sensitive genes in breast cancer samples from TCGA. PLoS One. Public Library of Science; 2019;14: e0210910.

Most importantly, a most heartfelt thank you to my friends and family. Technology becomes outmoded, research outdated, and ideas forgotten; but the impact we have through our relationships and camaraderie change the world we live in for the better. Thank you all for being a part of it and the motivation for me to do more.

Bobak David Kechavarzi

DEVELOPING BOTTOM-UP, INTEGRATED OMICS METHODOLOGIES FOR BIG DATA BIOMARKER DISCOVERY

The availability of highly-distributed computing compliments the proliferation of next generation sequencing (NGS) and genome-wide association studies (GWAS) datasets. These data sets are often complex, poorly annotated or require complex domain knowledge to sensibly manage. These novel datasets provide a rare, multi-dimensional omics (proteomics, transcriptomics, and genomics) view of a single sample or patient.

Previously, biologists assumed a strict adherence to the central dogma: replication, transcription and translation. Recent studies in genomics and proteomics emphasize that this is not the case. We must employ big-data methodologies to not only understand the biogenesis of these molecules, but also their disruption in disease states. The Cancer Genome Atlas (TCGA) provides high-dimensional patient data and illustrates the trends that occur in expression profiles and their alteration in many complex disease states.

I will ultimately create a bottom-up multi-omics approach to observe biological systems using big data techniques. I hypothesize that big data and systems biology approaches can be applied to public datasets to identify important subsets of genes in cancer phenotypes. By exploring these signatures, we can better understand the role of amplification and transcript alterations in cancer.

Huanmei Wu, PhD, Chair

TABLE OF CONTENTS

List of Tables x
List of Figures xi
List of Abbreviations xii
1Introduction 1
1.1 Objectives 1
1.1.1 Dissecting the expression landscape of RNA-binding proteins in human cancers 1
1.1.2 Bottom-up, integrated -omics analysis identifies broadly dosage-sensitive genes in breast cancer samples from TCGA 2
1.1.3 Deep Learning and transcriptional signatures for identifying key differentiating genes in cancer histologies 2
1.2 Significance 4
1.3 Contribution 5
1.4 Challenges 6
1.5 Organization of the study 7
2 Dissecting the expression landscape of RNA-binding proteins in human cancers 10
2.1 Background 10
2.2 Materials and methods 13
2.2.1 Data for healthy expression of RNA-binding proteins in 16 human tissues 13
2.2.2 Data for cancer expression of RNA-binding proteins for nine cancers in humans 14
2.2.3 Profiling for dysregulation of RNA-binding proteins and identification of strongly upregulated RNA-binding proteins across human cancers 14
2.2.4 Network and interaction properties of dysregulated RNA-binding protein in human cancers 15
2.2.5 Determination of prognostic impact of RNA-binding proteins for breast cancer 16
2.3 Results and discussion 17
2.3.1 RNA-binding proteins show significantly higher expression than non-RNA-binding proteins and other regulatory factors for 16 human tissues 17
2.3.2 RNA-binding proteins are dysregulated across cancers and a subset are strongly upregulated across a majority of cancers 22
2.3.3 Strongly upregulated and non-strongly upregulated RNA-binding Proteins exhibit significantly different within-group path lengths and variability in expression is related to the number of interactions 33
2.3.4 Survival contributions of RNA-binding proteins in breast cancer is related to network proximity to strongly upregulated RBPs and variability in expression across patients 37
2.4 Conclusions 42
3 Bottom-up, integrated -omics analysis identifies broadly dosage-sensitive genes in breast cancer samples from TCGA 45
3.1 Background 45
3.2 Materials and Methods 48

3.2.1 The Cancer Genome Atlas (TCGA) 48
3.2.2 Clinical Proteomic Tumor Analysis Consortium (CPTAC) 49
3.2.3 Gene amplification and deletion 49
3.2.4 Cancer gene profiling 50
3.2.5 Achilles shRNA 51
3.3 Results and discussion 51
3.3.1 Genomics and protein analysis 51
3.3.2 mRNA and protein dysregulation relative to copy number variation 54
3.3.3 Cancer gene profiling identifies broadly dosage-sensitive genes (BDSGs) 56
3.3.4 shRNA data defines the role of BDSGs in cancer cell line growth 58
3.4 Conclusions 61
4 Deep Learning and transcriptional signatures for identifying key differentiating genes in cancer histologies 63
4.1 Background 63
4.2 Methods 64
4.2.1 Data selection 64
4.2.2 Data preparation 65
4.2.3 Machine Learning Models 65
4.2.4 Feature selection and importance 67
4.3 Results and Discussion 68
4.3.1 -omics signatures in differing cancer histologies 68
4.3.2 Machine learning approaches 73
4.3.3 Feature importances in deep learning 74
4.4 Conclusions 78
5 Final Conclusions 79
6 Appendix 82
Supplemental Table 1 Ovarian cancer identified BDSGs, their chromosomal position, and description 82
Supplemental Figure 1 Heat map and hierarchical clustering of SHAP values for all gene deep neural network 88
7 References 89
Curriculum Vitae

LIST OF TABLES

Table 1: Strongly upregulated RNA-binding proteins identified from nine cancers in humans and their cancer relevant references. 27
Table 2: Broadly Dosage-Sensitive Genes (BDSGs).....57
Supplemental Table 1: Ovarian cancer identified BDSGs, their chromosomal position, and description. 82

LIST OF FIGURES

Figure 1: Representation of different automated learning methodologies. 3
Figure 2: Aims of the study and their relationships. 8
Figure 3: Flow chart showing the different steps in the analysis of expression levels of RNA-binding proteins for human cancers. 12
Figure 4: Expression levels of RNA-binding proteins (RBPs), non-RBPs, lncRNAs, miRNAs and transcription factors (TFs) for 16 human tissues. 20
Figure 5: Comparison of expression levels of RNA-binding proteins and non-RNA-binding proteins for 16 tissues from 80 healthy individuals studied in the Human BodyMap project. 21
Figure 6: Correlation matrix of overall log-ratio expression of RBPs across nine cancers. 23
Figure 7: Log-ratio of expression for cancer to healthy expression for RNA-binding proteins in nine human cancers. 26
Figure 8: Comparison of normalized network metrics (closeness, betweenness and degree) between strongly upregulated (SUR) and non-strongly upregulated (non-SUR) RNA-binding proteins. 32
Figure 9: Interaction profiles of RBPs. 34
Figure 10: Survival of patients with breast cancer for different expression levels and path lengths for within and between expression groups of RNA-binding proteins. 38
Figure 11: Comparison and distribution of prognostic impact based on expression dysregulation and expression variability in breast tissue. 41
Figure 12: Bottom-up, integrated analysis workflow. 48
Figure 13: Correlation coefficient cutoff. 51
Figure 14: Tumor-matched normal and tumor sample expression distributions. 52
Figure 15: mRNA log2 fold change from normal median to cancer. 53
Figure 16: mRNA fold change versus protein fold change. 54
Figure 17: Correlation coefficient distribution. 55
Figure 18: Protein vs mRNA fold changes with CNV amplification. 58
Figure 19: Heatmap and hierarchical clustering of shRNA knockdown. 60
Figure 20: Deep learning architecture for TCGA sample differentiation. 67
Figure 21: Correlation distributions for genomic features across BRCA and OV genes. 69
Figure 22: Expression distribution comparison between OV-, BRCA-, and non-BDSGs across genomic datatypes. 72
Figure 23: Deep learning test set prediction confusion matrix. 74
Figure 24: SHAP value distribution comparison between OV-, BRCA-, and non-BDSGs across genomic datatypes. 76
Supplemental Figure 1: Heat map and hierarchical clustering of SHAP values for all gene deep neural network. 88

LIST OF ABBREVIATIONS

BDSG: Broadly doseage sensitive genes

BRCA: Breast adenocarcinoma

CPTAC: Clinical Proteomics Tumor Analysis Consortium

CLIP: cross-linking and immunoprecipitation

DNN: Deep Neural Network

HBM: Human BodyMap

KM: Kaplan–Meier

KS test: Kolmogorov–Smirnov test.

lncRNA: long non-coding RNA

MAD: median absolute deviation

miRNA: microRNA

OV: Ovarian carcinoma

PAR-CLIP: photoactivatable-ribonucleoside-enhanced CLIP

PHD: plant homeodomain

PPI: protein–protein interaction

RBP: RNA-binding protein

RNA-seq: RNA sequencing

RNP: ribonucleoprotein

RPKM: reads per kilobase per millions of reads

SHAP: Shapely additive explanations

SUR: strongly upregulated

TCGA: The Cancer Genome Atlas

TF: transcription factor

TNF: tumor necrosis factor

1 **Introduction**

1.1 Objectives

The goal of this dissertation is to detail the transcriptional landscape of cancers and the alterations that occur from healthy to diseased states. This necessitates a multifaceted approach to gain a fundamental understanding of transcriptional regulation, how transcriptional abundancies and copy number alterations impact protein production, and finally, how transcriptional consistently can uniquely identify diseases.

1.1.1 Dissecting the expression landscape of RNA-binding proteins in human cancers

The expression signature of four transcriptional and posttranscriptional regulators (RNA binding proteins, micro-RNA, transcription factors, and long, non-coding RNA) were analyzed in 16 healthy human tissues. In particular, I focus on RNA binding proteins due to their importance in complex diseases and a poor understanding of their expression in human tissues. Based on these signatures I inferred their regulatory importance based on expression levels in key developmental tissues. Furthermore, I leveraged the publicly available data in The Cancer Genome Atlas to determine how their expression was altered from the healthy to disease state across nine human cancers. Examining these changes highlighted a subset of these regulators that are disrupted in cancer, suggesting potential biomarkers. Using additional interaction data, I estimated what downstream components might be impacted by these genes. This portion of research provides the groundwork in processing expression and metadata from TCGA, as well as profiling transcriptional activities of key expression regulators.

1.1.2 Bottom-up, integrated -omics analysis identifies broadly dosage-sensitive genes in breast cancer samples from TCGA

Now that I have profiled and identified aggressive patterns of enhanced expression, I have to understand how that impacts the actual production of protein within the cell. Often mRNA or transcript abundances are used as surrogates of the products produced by genes; however, new consortia offer proteomics data from processed TCGA samples. In this effort, copy number variation, mRNA foldchanges, and protein foldchanges for breast adenocarcinoma samples from TCGA were leveraged to globally determine dosage sensitivity. A subset of genes (broadly dosage sensitive genes, BDSGs) were identified by leveraging their correlation of copy number, mRNA, and protein features among samples. These genes can emphasize an important subclass of genes that could be useful for better understanding of cancer. By observing their behavior in single-cell shRNA models, one can attempt to estimate if these genes can contribute in the form of cancer suppression, cancer enhancement, or more generally as housekeeping genes.

1.1.3 Deep Learning and transcriptional signatures for identifying key differentiating genes in cancer histologies

My previous exploration emphasized a unique subset of genes that demonstrated a strong correlation across the three genomic features in breast cancer. Next, the analysis can be expanded to incorporate more recently available data and compare signatures between breast adenocarcinoma and ovarian carcinoma. Expanding the data allows us to observe any unique or overlapping genes between the two histologies. It also provides the opportunity to compare overall transcriptional differences between the two cancer

types. Deep neural networks were employed to train a model to differentiate samples between ovarian and breast cancer types. Utilizing deep learning afforded us the ability to use multimodal data to determine both the importance of the different genomic data types and the individual contributions on each of the genes in the classification task. Deep learning implementations allowed us to conduct these observations without having to perform feature extraction or prioritization prior to training (Figure 1). Shapley additive explanations allowed us to determine a gene's contribution in the effective sample classification. These values can provide a "weighting" to the previously identified BDSGs to determine if they do, in fact, differentiate sample types. Doing so helped illustrate if these genes are unique between histologies and identify which disease state a sample is in.

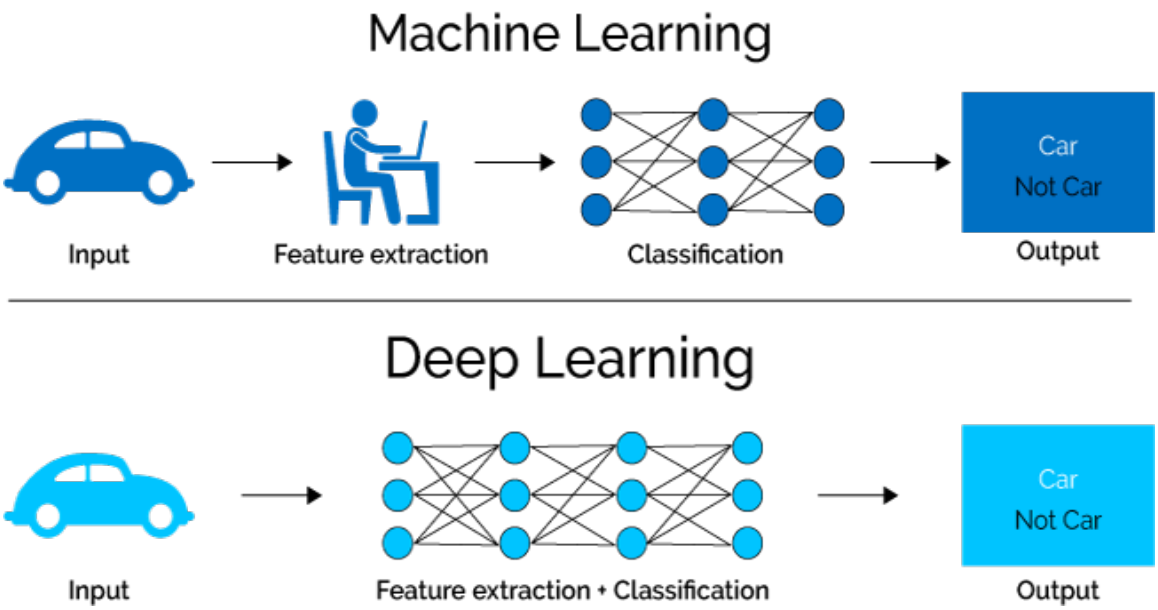


Figure 1: Representation of different automated learning methodologies. An illustrative example of the differences between machine learning and deep learning. In this case, it is not required to predetermine or prioritize features manually or through

personal intervention. Deep learning allows the feature extraction and classification steps to occur simultaneously.

1.2 Significance

Recently, DNA sequencing technologies have advanced tremendously. They can provide insights into structural variations as well as the abundance of transcripts within the samples, which benefit many biomedical studies. The proliferation and availability of these datasets also provide the opportunity to do largescale, cross-domain analyses. This is thwarted by finding datasets that have complete coverage over proteomic, genomic, epigenetic, and transcriptomic features. The Cancer Genome Atlas (TCGA) collects data for patients over 30 types of cancer and a wide array of -omics data. In 2015 approximately 40,000 women died of breast cancer. Understanding this disease space impacts a large population of suffering individuals and is aided by large, multidimensional datasets available for many patients.

It is essential to gain a firm grasp on the expression of genes in various tissues, how their signatures may change in each site, and the alteration that may occur from a healthy to a disease state. Previously, these systems-level observations were made with specific gene sets or pathways in mind. I leveraged big-data approaches to integrate large scale genomics, proteomics, and clinical data to observe overall trends throughout the genome.

- First, I observed how mRNA abundance levels for different classes of post-transcriptional and transcriptional regulators change from tissue to tissue, emphasizing their regulatory capacity in those sites.

- I extended the understanding by detailing their abundance changes in cancer; emphasizing potential dysregulators or onco-genes.
- I integrated proteomics and copy number variation data to understand how the genes adhere to the central dogma and if the alterations impact the production of proteins.
- Finally, I generated comprehensive, integrated views across multiple histologies to determine perturbations that may emphasize putative biomarkers.

Ultimately, I generated a big data omics system to better perform bottom-up hypothesis testing. It will look at the sample data in their entirety to identify meaningful signatures that can elucidate new biomarkers that would have been otherwise missed due to a bias.

1.3 Contribution

I engineered a computational framework to perform big-data queries across varied, complex biological datasets. The investigation also provided a technical solution in data structures and workflow methodologies for large-scale genomics. A signature catalog of gene expression over thousands of samples, tissues, and cancer types was produced for large-scale analytics. The integration of protein fold changes, mRNA abundances, and copy number variation across a variety of cancer types and patients is an unprecedented opportunity to conduct truly systems-level observations. Leveraging deep learning methodologies further emphasized the disruption occurring in the tissues and how seemingly unrelated genes are acting to differentiate samples. Furthermore, a data dashboard of the results was created and distributed using container-based virtualization to preserve the results. This effort itself demonstrates the strength of non-traditional data

storage and processing, such as in-memory databases and deep neural networks in extending systems biology analyses.

1.4 Challenges

One of the most pressing challenges in an effort such as this was the missing or incomplete data. Aim 1 began early in the development life of the TCGA project and sample coverage across the multiple histologies was only just beginning. Some data had to be discarded because of incompleteness or low sample diversity. In Aims 2 and 3, for example, samples had to be excluded because of missing proteomics data. Careful consideration had to be given early on to data design and engineering. Schema needed to be resilient enough to add or remove data, but not become too unwieldy for the later, intensive computational tasks. For machine learning tasks, considerable work was necessary to define negative controls. Analyzing such large volumes of data often introduce correlations that can add false signals for sample differentiation. For example, for Aim 3 several models were able to differentiate samples because the model identified age-differentiating genes that illustrated an age bias between ovarian and breast cancer samples.

Careful unit testing and best software practices for parallelized systems were used to quickly perform the computations identifying correlational patterns across three genomic data types and thousands of samples in Aim 2. In general, the data volume and computational complexity derived most the challenges in these analyses, and even directed the research to leverage deep-learning methodologies. Most standard machine learning or clustering approaches do not perform well with several hundred-thousand features. To do so would have required feature reduction techniques and would have

defeated the purpose of performing these bottom-up analyses. Even still, training the deep models required careful consideration in the architecture because of the large volume of features, even GPUs can only accommodate a certain volume of data.

These challenges helped to emphasize how technological consideration and data engineering have become integral in scientific study design. Addressing these aspects encouraged developing skills and solutions in high performance databases and distributed systems, as well as the aforementioned deep learning frameworks such as TensorFlow and Keras.

1.5 Organization of the study

Public transcriptomic datasets are utilized to describe the activity of various classes of genes and determine if there is an impact or correlation to a disease state. Figure 2 illustrates the aims and their relationships. Figures 3 and 12 illustrate the workflows for the RNA binding protein expression exploration and BDSG identification, respectively. In general, I leverage an aggregated normal control to determine expression changes from healthy to disease states. Afterwards additional analytical steps are performed to determine impacts on survivability or importance to the disease phenotype.

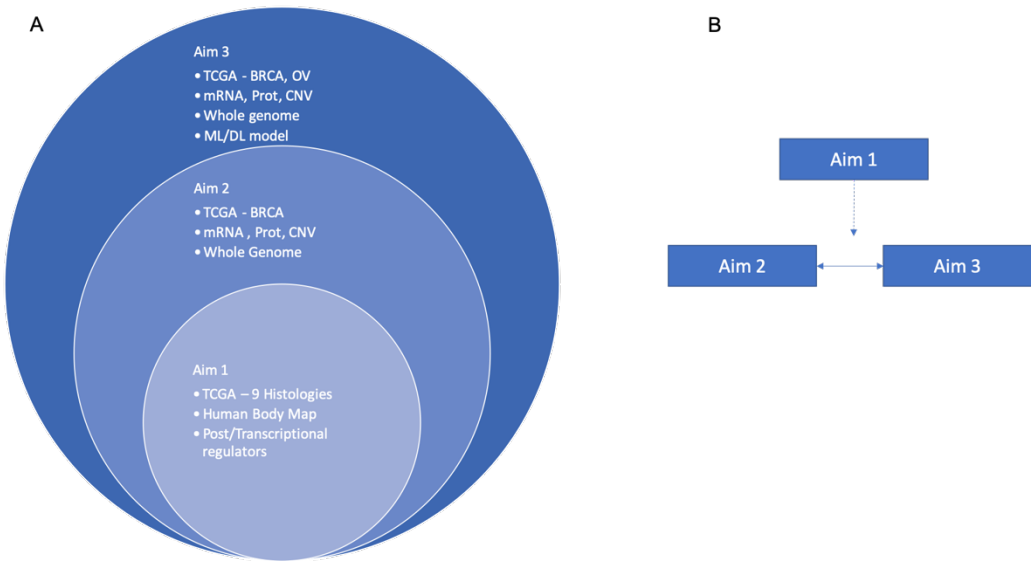


Figure 2: Aims of the study and their relationships. (A) Aim 1 begins the study with the most comprehensive assessment across multiple tissues, but only regarding one genomic data type. Aim 2 expands the investigation of breast adenocarcinoma by including two additional genomic datatypes. Aim 3 then looks across multiple cancer types using the same methodology as Aim 2 to derive a feature set for deep neural network models for sample identification. (B) Methodological relationship of the aims.

The first aim of the dissertation provides an initial overview of the expression of cornerstone transcriptional regulators in a variety of healthy human tissues. The study expands to include the changes in their expression from the healthy to the cancerous state. In particular, RBPs are identified which were substantially enhanced and had impacts on patient survivability. The second goal allowed us to expand the understanding of past mRNA changes. With the inclusion of copy number and protein data, genes are identified whose genomic alterations are reflected in their protein production. Additionally, these genes behavior are described via cell line models and shRNA experiments. Finally, the differences in these gene sets are compared across

cancer types and can build predictive models to demonstrate how these genes differentiate samples.

2. Dissecting the expression landscape of RNA-binding proteins in human cancers

2.1 Background

RNA-binding proteins (RBPs) have been identified as key regulatory components interacting with the RNA within a cell. Their function is largely dependent on their expression and localization within a cell. They may be involved in processes ranging from alternative splicing to RNA degradation. Combining together, RBPs form dynamic ribonucleoprotein (RNP) complexes, often in a highly combinatorial fashion that can affect all aspects of the life of RNA [1–3]. Due to their central role in controlling gene expression at the post-transcriptional level, alterations in expression or mutations in either RBPs or their binding sites in target transcripts have been reported to be the cause of several human diseases such as muscular atrophies, neurological disorders and cancer (reviewed in [4–7]). These studies suggest there is precise regulation of expression levels of RBPs in a cell.

A recent system-wide study of the dynamic expression properties of yeast RBPs showed that RBPs with a high number of RNA targets are likely to be tightly regulated. Significant changes in their expression levels can bring about large-scale changes in the post-transcriptional regulatory networks controlled by them [8]. RBPs have also been shown to autoregulate their expression levels. Fluctuations in the expression of autoregulatory RBPs are significantly decreased [9]. These results show that a low degree of expression noise for RBPs is a characteristic feature of their normal state.

Cancer is a complex genetic disease and many of its regulatory factors have been identified as being irregularly expressed. In particular, changes in the normal expression

of RBPs have been shown to alter their function leading to a cancer phenotype [10]. Enhanced eIF4E and HuR expression levels have been implicated in initiating translation of mRNAs encoding mostly for pro-oncogenic proteins and other cancer-promoting processes. For instance, Sam68 regulates the alternative splicing of cancer-related mRNAs [10]. Yet another example is the cell-specific alternative splicing of FAS (Fas cell surface death receptor, a member of the TNF receptor superfamily) mRNA. This has been linked to cancer predisposition depending on whether the pro- or anti-apoptotic protein form is produced as a result of the interplay between various RBPs on the FAS transcript [11–14]. In some cases, disruption of the functionality of RBPs, although without directly acting on oncogenic genes, has been shown to affect alternative splicing regulation or the regulation of alternative cleavage mechanisms on transcripts, which can lead to the development of cancer [15,16].

In a recent study, Castello and co-workers [17] utilized cross-linking and immunoprecipitation (CLIP) and photoactivatable-ribonucleoside-enhanced CLIP (PAR-CLIP) to isolate and validate, via proteomics, a set of approximately 850 high-confidence RBPs in humans. These approaches can be used to catalogue and study RBPs and their post-transcriptional networks in healthy and diseased states. By knowing the low degree of expression variation that is tolerated by RBPs in a healthy state and identifying them in mammalian systems, we can begin to investigate their dysregulation profiles in various disease conditions.

In this study, I analyzed the expression patterns of RBPs in a set of 16 healthy human tissues and compared their fold change in expression levels in nine human cancers using the high-resolution expression profiles based on RNA sequencing (RNA-seq)

available from the Human BodyMap (HBM) [18]and the Cancer Genome Atlas (TCGA) [19] (see Figure 3, which outlines the different steps, and Materials and methods). The network properties of a set of 31 RBPs were compared, which were found to be strongly upregulated (SUR) for most of the cancers studied. The network properties may help to determine the cause of the altered expression for the RBPs. Finally, a subset of RBPs was identified based on their expression profiles and network metrics and their contribution to the survival of patients with breast cancer was investigated.

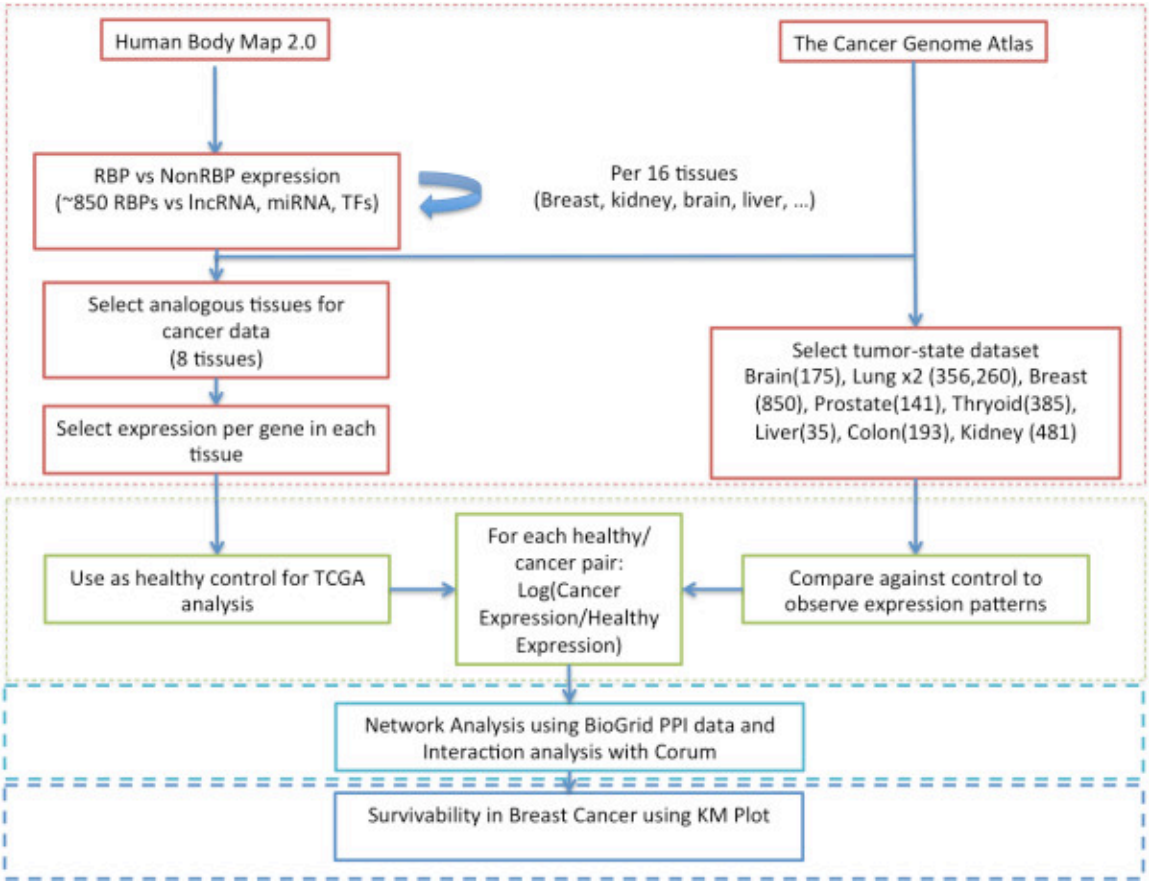


Figure 3: Flow chart showing the different steps in the analysis of expression levels of RNA-binding proteins for human cancers. The flow chart shows the acquisition and preparation of data (red), determination of patterns of dysregulation (green), network and interaction analysis (light blue), and survival analysis (dark blue).

2.2 Materials and methods

2.2.1 Data for healthy expression of RNA-binding proteins in 16 human tissues

The general workflow is illustrated in Figure 3. RNA-seq data for 16 different human tissues from ArrayExpress [20] (Accession no. E-MTAB-513), which is part of the Human BodyMap (HBM) 2.0 project [18,21], was obtained for expression profiling. This data represents the healthy RNA transcript levels of male and female individuals aged 19 to 86, for 16 tissues: adipose, adrenal, brain, breast, colon, heart, kidney, liver, lung, lymph node, ovary, prostate, skeletal muscle, testes, thyroid and white blood cells. Expression data from the HBM project was quantified per transcript using the current annotations of the human genome from the Ensembl. This is available as reads per kilobase per millions of reads (RPKM) for each sample and hence can be compared across and within tissues. Therefore, each of the 16 tissues has a single RPKM value for the expression level of each transcript. A total of 850 genes experimentally characterized as RBPs in the human genome were obtained from a previous publication [17] and 4,647 transcripts associated with these RBPs were identified within the HBM set. The remaining set of 102,462 transcripts were classified as non-RBPs in this study. To examine the other regulatory factors in humans I obtained a set of 9,440 long non-coding RNAs (lncRNAs) from a Gencode study [18,22], 529 microRNAs (miRNAs) from miRBase [23] and 1,231 transcription factors (TFs) from the DBD database [24]. For each of the 16 tissues distribution of the RPKM values for transcripts associated with RBPs and non-RBPs were compared, as well as the distribution of expression levels of transcripts associated with RBPs with other regulatory factors to study their relative effect on regulatory control at the tissue level.

2.2.2 Data for cancer expression of RNA-binding proteins for nine cancers in humans

The cancer expression data was downloaded from TCGA [19]. TCGA provides multi-level data (clinical, genome sequencing, microarray, RNA sequencing and so on) procured from a number of institutions, from a variety of patients, for over 25 cancers. In this study, RNAseq V2.0 data for 2,876 patients were collected spanning nine cancers analogous to eight of the tissues in the HBM dataset: breast (850 patients), brain (175 patients), colon (193 patients), kidney (481 patients), liver (35 patients), two for lung (356 and 260 patients), prostate (141 patients), and thyroid (385 patients). For each cancer I collected the expression levels for each gene for all patients and determined a median representative level and MAD. This defines the genes' RNA expression levels and variability in the relevant cancer state. Likewise, cancer expression and variation were determined for the group of non-RBP genes from HBM as a complementary group for later network, interaction, and expression analyses. Hierarchical clustering of RBP expression for these nine cancers was performed in R, to determine if similar cancers and tissues group together (Figure 6). Clustering results verified that the collected and amalgamated data are an accurate representation of their anatomical origin and can be utilized to draw further conclusions.

2.2.3 Profiling for dysregulation of RNA-binding proteins and identification of strongly upregulated RNA-binding proteins across human cancers

For each gene identified as an RBP, I calculated a median expression level of its transcript products in the HBM data when there were multiple protein coding transcripts. To determine the extent of dysregulation in RBPs across cancers, for each cancer the log-ratio of the median expression in the cancer state over its expression in the associated

healthy state was calculated. This allowed us to determine for the nine cancers if a particular gene annotated as an RBP is upregulated, downregulated or does not change in expression level in cancer states. Based on this analysis, if an RBP has a log-ratio of expression level greater than 9 across six or more of the studied cancers, it was classified as being SUR. Otherwise, it was categorized as non-SUR. I focused mainly on defining characteristics unique to these SUR RBPs that differentiate them from other RBPs and non-RBPs. SUR genes as defined here were also observed in non-RBPs and a hypergeometric test was performed to examine potential differences in the proportionality of SUR RBPs and non-SUR RBPs between the two functional classes. The genes associated with RBPs and non-RBPs were also classified by their level of expression variability in a cancer, measured as the MAD value of the fold change in expression for the profiled patients for the cancer. If a gene's variability within a cancer was above the 75th percentile, it was considered highly variable, below the 25th percentile it was considered least variable and the remainder were considered moderately variable.

2.2.4 Network and interaction properties of dysregulated RNA-binding protein in human cancers

The most recent BioGRID [25] protein–protein interaction (PPI) information (version 3.2.97) was downloaded and used to construct an undirected network of interactions documented in humans. These interactions were used to determine if there were any differences in network properties between the two classifications of dysregulated RBPs, that is, SUR and non-SUR RBPs. This allowed the determination of the potential importance of the classifications for these RBPs. For example, if an SUR RBP forms a hub, it could cause patterns of dysregulation in other, associated interactors.

Network centrality measures were compared such as degree, closeness and betweenness as well as clustering coefficients and shortest paths between nodes, for different RBP classes utilizing the R package igraph [26]. For shortest paths, I calculated the mean shortest paths for a SUR RBP to other SUR RBPs and SUR RBPs to non-SUR RBPs. The overall average path length was obtained between each RBP/non-RBP and SUR RBP/non-SUR RBP combination.

Manually curated experimentally characterized human protein complex data was obtained from CORUM [27], to determine the general promiscuity of RBPs in forming complexes. Then 5,217 protein complexes were mapped to the RBPs. For SUR RBPs and non-SUR RBPs the frequency of membership in CORUM complexes were calculated, as well as the mean complex size. This information together with the log-ratios of expression levels between healthy and cancer states in the tissues, allowed us to address whether SUR RBPs are enriched in protein complexes and/or occur in larger or smaller complexes. This analysis also allowed us to test the relation between the extent of an RBP's dysregulation in the context of its membership.

2.2.5 Determination of prognostic impact of RNA-binding proteins for breast cancer

A gene's prognostic impact is the gene's ability to impact positively or negatively patient survival. The prognostic impact for each gene was determined using data from the Kaplan–Meier (KM)-Plotter [28], which was determined from microarray experiments for over 20,000 genes for 1,800 breast cancer patients. For each gene in the RBP and non-RBP groups, they were further categorized as SUR or non-SUR and high or low variability in expression. The significance [-log(KM-plotter *P*)] of the prognostic impacts were compared within and between these groups.

Based on the network analyses, the genes were ranked in descending order based on their mean path lengths to the classification of dysregulated genes (SUR vs non-SUR). Path length calculations were determined from a distance matrix generated by the network analysis. From the ranked list of genes, five genes were selected with the shortest and longest mean path lengths and took a random sample of five genes with intermediate mean path lengths. This provided information on the prognostic impact associated with increased gene expression.

2.3 Results and discussion

2.3.1 RNA-binding proteins show significantly higher expression than non-RNA-binding proteins and other regulatory factors for 16 human tissues

In eukaryotes, transcription and translation occur in different compartments. This gives a plethora of options for controlling RNA at the post-transcriptional level, including splicing, polyadenylation, transport, mRNA stability, localization and translational control [1,2]. Although some early studies revealed the involvement of RBPs in the transport of mRNA from the nucleus to the translation site, increasing evidence now suggests that RBPs regulate almost all of these post-transcriptional steps [1–3,29]. RBPs have a central role in controlling gene expression at the post-transcriptional level. Alterations in expression and mutations in either RBPs or their RNA targets (the transcripts that physically associate with the RBP) have been reported to be the cause of several human diseases, such as muscular atrophies, neurological disorders and cancer [4,5,7,30].

Therefore, I first chose to study the mRNA expression levels of a repertoire of approximately 850 experimentally determined RBPs for all 16 human tissues for which

expression data are available from the Human BodyMap 2.0 Project [18,21](see Materials and methods). This analysis clearly showed that RBPs are significantly more highly expressed ($P < 2 \times 10^{-16}$, Wilcoxon test) than non-RBPs in all of the tissues (Figure 5). Closer inspection of the trends also revealed that some tissues, such as those from the testes, lymph and ovary, had particularly high RBP expression compared to non-RBPs. To determine the regulatory effect of RBPs at the post-transcriptional level compared to other regulatory factors, such as transcription factors (TFs), microRNAs (miRNAs) and long non-coding RNAs (lncRNAs), their expression levels were compared for different human tissues (see Figure 4 and Materials and Methods).

This analysis further revealed that the expression levels of RBPs are significantly different for these 16 tissues compared to these families of regulatory factors ($P < 2 \times 10^{-16}$, Kruskal–Wallis test). Further analysis to compare the expression levels of RBPs and TFs across tissues revealed that except for the heart, kidney, ovary and testis, RBPs are significantly more highly expressed than TFs ($P < 0.05$, Wilcoxon test). These observations suggest that in most tissues, the magnitude of expression of RBPs is more prominent than even TFs, possibly indicating their central role in controlling gene expression than previously anticipated. The observation that RBPs are not significantly more highly expressed than TFs in heart, kidney and gonadal tissues like the testis and ovary suggests that both transcriptional and post-transcriptional regulators are equally important in terms of their expression levels in these tissues. In contrast, tissues like the liver ($P < 3.57 \times 10^{-11}$, Wilcoxon test) and white blood cells ($P < 3.85 \times 10^{-5}$, Wilcoxon test) were found to have significantly higher expression for RBPs compared to TFs,

possibly indicating the importance of post-transcriptional regulation in the regenerative capabilities of a tissue or in monitoring inflammation and immune response.

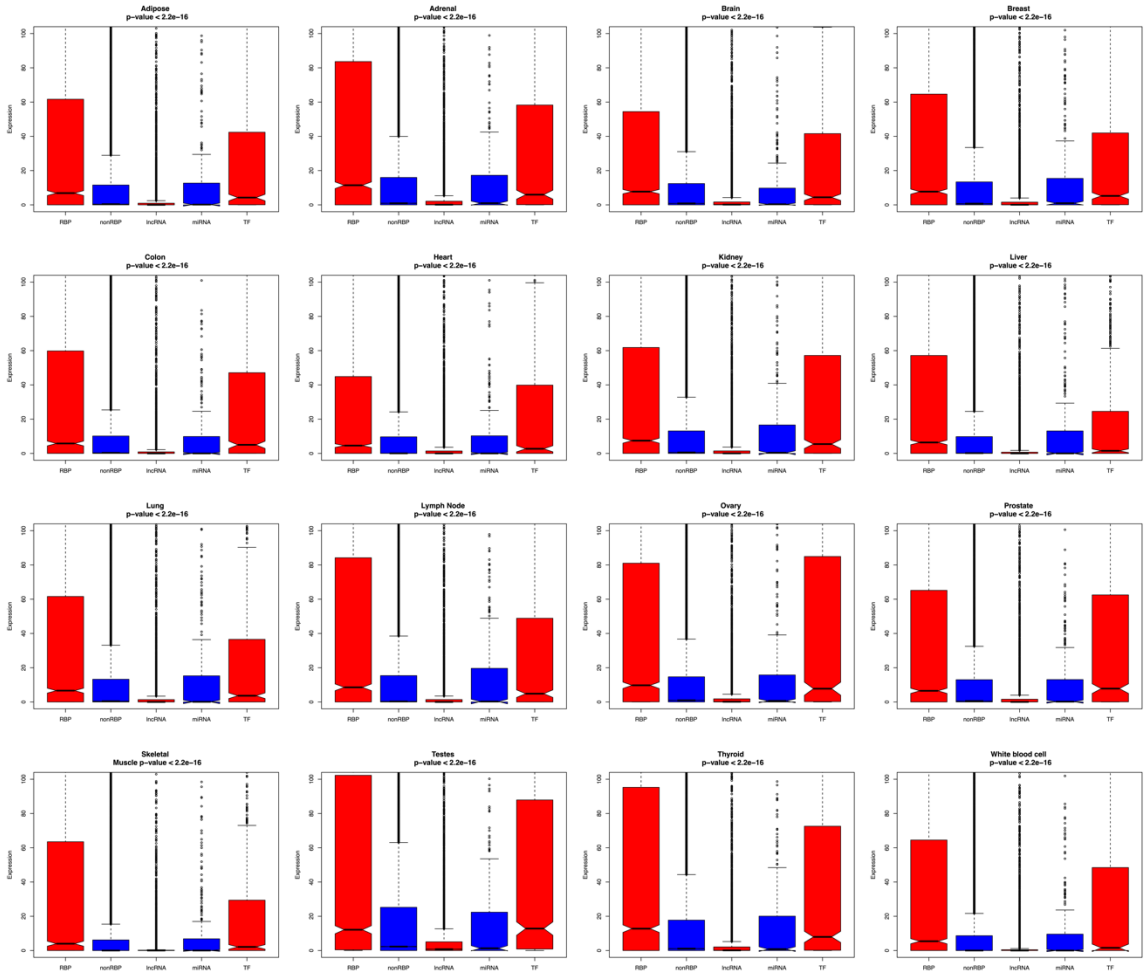


Figure 4: Expression levels of RNA-binding proteins (RBPs), non-RBPs, lncRNAs, miRNAs and transcription factors (TFs) for 16 human tissues. Each of the 16 plots illustrates the significant differences in expression levels of RBPs ($P < 2 \times 10^{-16}$, Wilcox test) for adipose, adrenal, brain, breast, colon, heart, kidney, liver, lung, lymph node, ovary, prostate, skeletal muscle, testes, thyroid and white blood cell tissues, compared to the other regulatory factors. The *x*-axis is the category of the observed factor and the *y*-axis is the expression level.

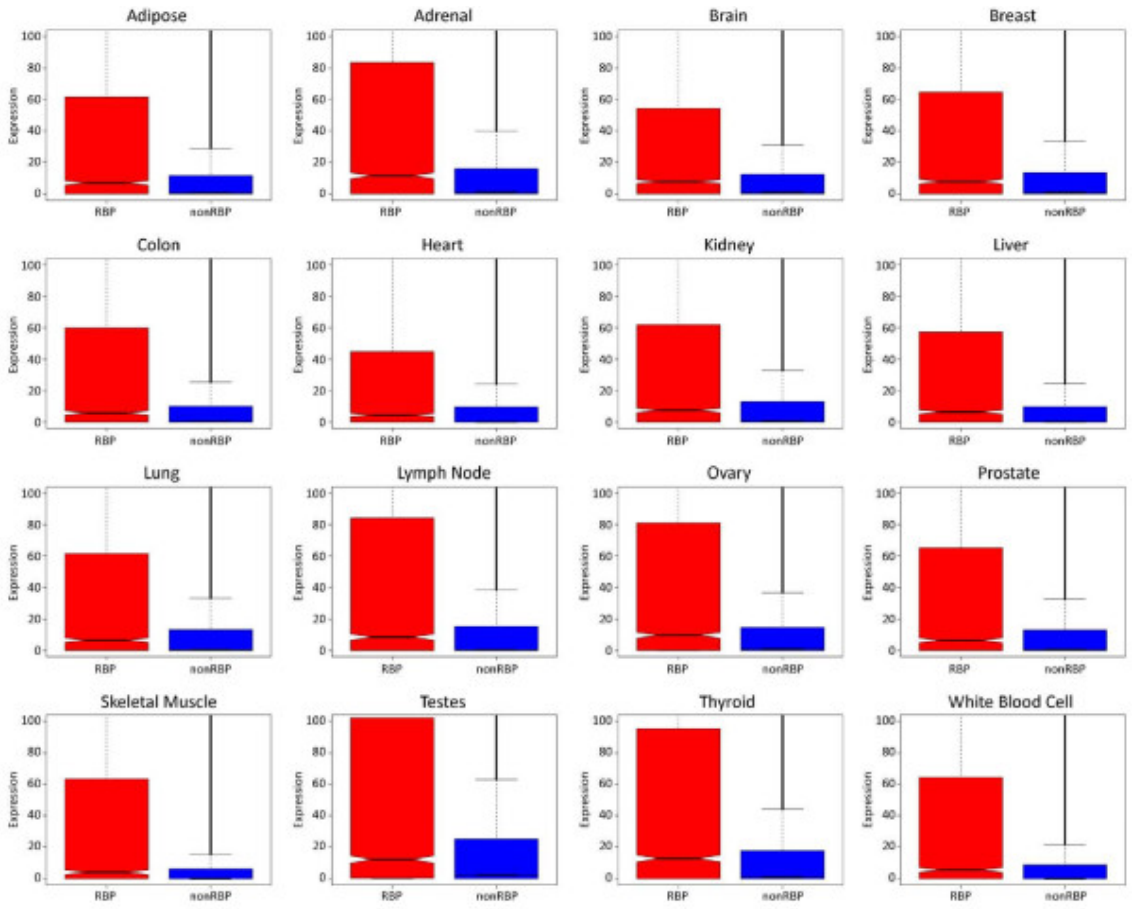


Figure 5: Comparison of expression levels of RNA-binding proteins and non-RNA-binding proteins for 16 tissues from 80 healthy individuals studied in the Human BodyMap project. Each of the 16 plots illustrates the significant differences in expression levels in RBPs ($P < 2 \times 10^{-16}$, Wilcoxon test) across adipose, adrenal, brain, breast, colon, heart, kidney, liver, lung, lymph node, ovary, prostate, skeletal muscle, testes, thyroid, and white blood cell tissues. The *x*-axis is the category of the observed factor and the *y*-axis is the expression level.

The fact that RBPs exhibit a particularly high level of expression in some tissues suggests a need for extensive post-transcriptional control of gene expression in them. For example, the coordinated and cyclic processes of spermatogenesis in testes necessitate the essential temporal and spatial expression of pertinent genes [31]. In the human prostate, slight alterations to the androgen receptor functionality [32] or transcription factors [33] have been shown to lead to a cancerous state. These trends suggest that a significant fraction of the RBPome might play an important regulatory role in diverse human tissues, although in some gonadal and developed tissues, RBPs and TFs had similar levels of expression. My results show that the high expression of RBPs is especially important in developmentally important tissues suggesting that any patterns of dysregulation could strongly effect these tissues [8].

2.3.2 RNA-binding proteins are dysregulated across cancers and a subset are strongly upregulated across a majority of cancers

Based on the understanding of the expression landscape of RBPs in healthy human tissues, I next asked whether RBPs are dysregulated across cancers (see Materials and methods). Since expression data for healthy tissue was available for eight tissues from the Human BodyMap project corresponding to a set of nine different cancers profiled in the Cancer Genome Atlas (TCGA), I calculated the log-ratio of expression levels of RBPs in the healthy to cancerous states in each of the nine cancers (Materials and methods). Positive values represent a shift towards upregulation, or, more generally, increased transcript abundance. Negative log-ratios represent a trend of downregulation or decreased abundance. The log-ratio expression profile matrix for the nine cancers was hierarchically clustered to show patterns of similar dysregulation (Figure 6). Cancers in

similar tissues (lung adenocarcinoma and lung squamous carcinoma) are clustered together suggesting a similar degree of dysregulation of the RBP repertoire. The analysis also revealed that similar cancers, such as adenocarcinomas were clustered together. These trends indicate that expression ratios are reliable for profiling cancers with unique morphologies in various body locations.

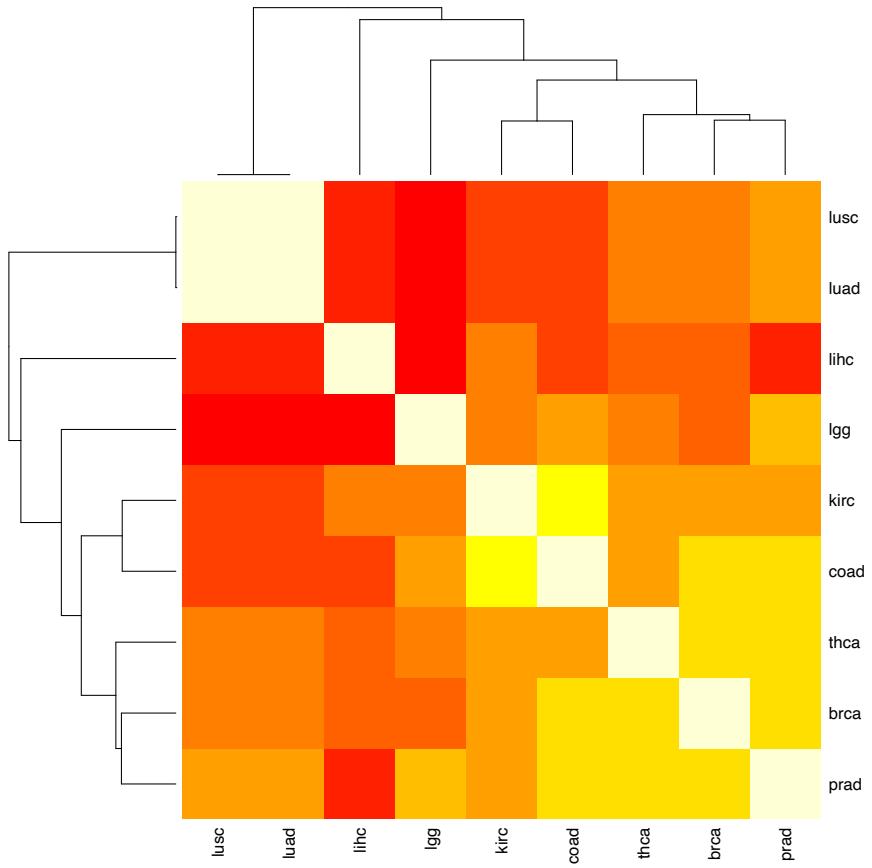


Figure 6: Correlation matrix of overall log-ratio expression of RBPs across nine cancers. The matrix shows the clustering of similar tissue sites and similar cancer types.

An analysis of the log-ratios representing the fold changes in expression of RBPs between healthy and cancerous states for nine different cancers allowed us to define a criterion for classifying RBPs as strongly upregulated (SUR) or not (non-SUR) (Figure 7, Materials and methods). If an RBP, across six of the nine cancers, was found to have a

log-ratio for expression level change of at least nine, it was classified as highly dysregulated, otherwise it was not considered to be a significantly dysregulated RBP. This also corresponded to the RBPs that belonged to the upper quartile of the fold changes in expression across cancers. According to this criterion, all the RBPs that had at least a ninefold change in expression were found to be only upregulated and hence this group was termed SUR RBPs (Figure 7). Table 1 lists these 31 SUR RBPs.

I then asked whether tumor-matched normal expression data for TCGA samples can further support the set of SUR RBPs identified here. Although 'normal' site tissue samples from TCGA cannot provide an adequate control, since these samples are collected from a cancerous tissue and it is entirely feasible that the expression levels would still be in a state of dysregulation at the neighboring sites, this analysis can still provide an additional level of support for SUR RBPs. Additionally, it is not possible to control for morphological types of tumors, which, depending on their type, can affect more than just the site of the tumor growth. Nevertheless, I profiled the tumor-matched normal expression levels that are available for eight of the nine cancer types with varying number of samples for breast (106 patients), colon (20 patients), kidney (69 patients), liver (49 patients), two types of lung cancers (57 and 50 patients), prostate (45 patients) and thyroid (58 patients). As suspected, the fold changes in expression for all the genes across eight cancers were minimal (median [IQR] 0.055 [-0.28:-0.39]), suggesting that tumor-matched normal expression data may not reflect a true healthy control. However, when compared to the fold changes in expression levels for RBPs and non-RBPs in the tumor-matched samples across cancers, RBPs exhibited significantly higher fold changes compared to non-RBPs (median [IQR] 0.104 [-0.07:0.29] for RBPs versus median [IQR]

-0.034 [-0.39:0.25] for non-RBPs, $P < 2.2 \times 10^{-16}$, Wilcoxon test) clearly indicating that RBPs are still significantly upregulated in tumors.

Further analysis to test for the enrichment of RBPs in the top quartile of upregulated genes across cancers revealed that RBPs are strongly over-represented in this list ($P = 1.62 \times 10^{-93}$, hypergeometric test). I also found that all the SUR RBPs are significantly dysregulated ($P < 0.001$, *t*-test comparing tumor and matched normal samples) in at least four of the eight cancers profiled. When strigency was raised to identify an RBP to be dysregulated in at least six or more cancer types, 24 of the original 31 SUR RBPs were detected at $P < 0.001$. Very few SUR RBPs from the cancer types KIRC and LIHC were found to be significantly altered in the tumor-matched analysis. While most of the SUR RBPs were found to be upregulated in the tumor-matched analysis, I also found cases of downregulation. Nevertheless, SUR RBPs as a group were also found to be strongly over-represented in the top quartile of the upregulated set in the tumor-matched analysis ($P = 2.16 \times 10^{-8}$, hypergeometric test), further supporting the notion that SUR RBPs identified using an external healthy control across a broad range of cancers are a confident set of dysregulated RBPs.

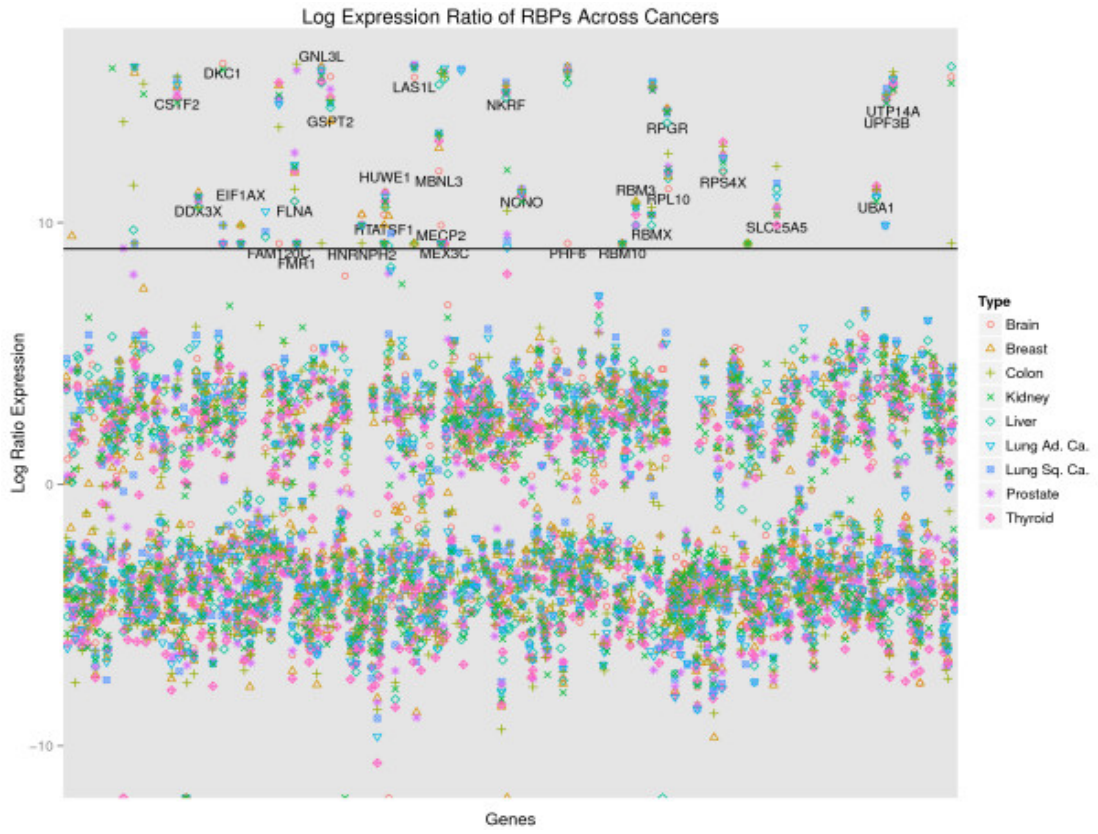


Figure 7: Log-ratio of expression for cancer to healthy expression for RNA-binding proteins in nine human cancers. The *x*-axis is an index of all the RNA-binding proteins that could be extracted from the expression data in the Cancer Genome Atlas. The *y*-axis is the ratio of the median expression level for each gene across patients versus the observed expression in the Human BodyMap data. Marked are the 31 strongly upregulated RBPs that have an expression ratio over nine across more than half of the studied cancers.

Associated gene name	Description	References
CCDC124	Coiled-coil domain containing 124	
CSTF2	Cleavage stimulation factor, 3′ pre-RNA, subunit 2, 64 kDa	[34]
DDX3X	DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, X-linked	[35–37]
DKC1	Dyskeratosis congenita 1, dyskerin	[38–40]
EIF1AX	Eukaryotic translation initiation factor 1A, X-linked	
FAM120C	Family with sequence similarity 120C	
FLNA	Filamin A, alpha	[41–44]
FMR1	Fragile X mental retardation 1	
GNL3L	Guanine nucleotide binding protein-like 3 (nucleolar)-like	[45,46]
GSPT2	G1 to S phase transition 2	
HNRNPH2	Heterogeneous nuclear ribonucleoprotein H2 (H')	
HTATSF1	HIV-1 Tat specific factor 1	
HUWE1	HECT, UBA and WWE domain containing 1, E3 ubiquitin protein ligase	[47]
LAS1L	LAS1-like (*Saccharomyces cerevisiae*)	
MBNL3	Muscleblind-like splicing regulator 3	
MECP2	Methyl CpG binding protein 2 (Rett syndrome)	
MEX3C	Mex-3 homolog C (*Caenorhabditis elegans*)	
NKRF	NFKB repressing factor	[48]
NONO	Non-POU domain containing, octamer-binding	[49,50]

PHF6	PHD finger protein 6	[51–53]
RBM10	RNA-binding motif protein 10	
RBM3	RNA-binding motif (RNP1, RRM) protein 3	[54–57]
RBMX	RNA-binding motif protein, X-linked	[58]
RBMX2	RNA-binding motif protein, X-linked 2	
RPGR	Retinitis pigmentosa GTPase regulator	
RPL10	Ribosomal protein L10	
RPS4X	Ribosomal protein S4, X-linked	
SLC25A5	Solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator), member 5	
UBA1	Ubiquitin-like modifier activating enzyme 1	[59,60]
UPF3B	UPF3 regulator of nonsense transcripts homolog B (yeast)	
UTP14A	UTP14, U3 small nucleolar ribonucleoprotein, homolog A (yeast)	

Table 1: **Strongly upregulated RNA-binding proteins identified from nine cancers in humans and their cancer relevant references**.

Non-RBP log-ratios showing the expression changes were also calculated using the external healthy data to determine if the proportion of strongly upregulated genes (SURs) in RBPs is significantly enriched. The proportions were significantly different ($P < 0.05$, hypergeometric test) with RBPs having a higher proportion of SURs than non-RBPs. Several of these SUR RBPs were annotated to function in important biological processes, such as regulation of gene expression, transcriptional regulation and transport of biomolecules, although very few studies have explored their role in the context of

post-transcriptional control, suggesting that their functional roles are far more diverse than previously understood and appreciated.

Of these RBPs classified as SUR RBPs, several have already been implicated in complex genetic disorders and cancer or in cellular regulation and proliferation. Identified RBPs, such as NONO, are involved in RNA biogenesis and DNA double-strand break repair, and have been found to be regulated by other factors, when dysregulated potentially promote carcinogenesis [49]. DDX3X, a member of the DEAD box RNA helicase family, has been shown to affect Wnt pathways, which leads to the developments of cancers [35]. DDX3X has also been demonstrated to promote growth and neoplastic transformation of breast epithelial cells [36]. Another SUR RBP, LAS1L was identified to interact with PELP1, which is implicated in pancreatic cancers [61]. HUWE1 is a member of the HECT family of E3 ubiquitin ligases, which has been identified as being overexpressed in breast, lung and colorectal cancers [62]. Indeed, increasing evidence now points to the role of novel ubiquitin-protein ligases in binding to RNA [63,64]. For instance, ubiquitin-like fold has been recently shown to be independently enriched in novel unconventional RBPs identified in the yeast genome [65]. The RNA-binding protein RBM3 is associated with cisplatin sensitivity, the probability of a patient becoming resistant to cisplatin treatment and a positive prognosis in epithelial ovarian cancer [54]. RBM3 has seldom been found expressed in normal tissues, but it is more expressed in common cancers, particularly for the nuclear expression of Estrogen-Receptor (ER) positive tumors. These findings suggest the possible utility of the gene as a positive prognostic marker [55,56].

PHF6 encodes a plant homeodomain (PHD) factor containing four nuclear localization signals and two imperfect PHD zinc-finger domains and it has been proposed that it has a role in controlling gene expression [66]. Inactivating mutations in PHF6 cause Börjeson-Forssman-Lehmann syndrome, a relatively uncommon type of X-linked familial syndromic mental retardation [66–68] . Recent studies show that mutations of this gene are implicated in the development of T-cell acute lymphoblastic leukemia and mutations have been detected in other forms of leukemia as well, suggesting a strong role in tumorigenesis [51,52]. For other nucleolar proteins such as dyskerin (DKC1), which is responsible for the biogenesis of ribonucleoproteins and telomerase stability, the loss or gain of functions is associated with tumorigenesis [38–40]. Filamin A (FLNA) is an actin-binding protein, which interacts with a number of proteins including signaling molecules and membrane receptors, and its expression has been correlated with metastases in prostate and lung cancers [41,42]. A recent study demonstrated the role of FLNA as a nucleolar protein that associates with the RNA polymerase I (Pol I) transcription machinery to suppress rRNA gene transcription [69]. Although further confirmation of how the global RNA-binding role of unconventional RBPs, like the E3 ubiquitin ligase HUWE1, contribute to cancer is needed, increasing evidence suggests that several enzymes and kinases bind to RNAs to control numerous cellular processes [57,63] [65,70]. Recent genome-wide screens for novel RBPs further support these observations, suggesting that unconventional RBPs are enriched for enzymatic functions [65,71]. Functional enrichment analysis of SUR RBPs using the DAVID functional annotation system [72] revealed that RNA splicing, nucleotide binding and ribosome biogenesis were the common biological processes associated with these proteins, with a

significant fraction of them associated with nucleolus and nuclear lumen cellular components.

My observations combined with the existing corpus of literature in support of the roles for several of these SUR RBPs in cancerous states, suggest that their dysregulation could be the cause or result of the cancer phenotypes, especially given that even slight alterations in the expression levels of RBPs can bring about large-scale changes in the RBP–RNA interaction networks that they control [8]. It is important to note that although some of these SUR genes shown in Table 1 have been described in relation to cancer, there is little evidence in support of their contribution to either being RBPs or their post-transcriptional network as a contributing factor for the cancer phenotype. The results in this study implicate them as a strongly upregulated set of RBPs across multiple cancers. My analysis also corroborates that these significantly dysregulated RBPs are not an artifact of aberrations in calculations, or due to variability in patient expression data mainly because: (1) most of the patient sample sets are at least of the order of 100 for the cancers studied and (2) fold changes in expression levels between healthy and cancerous states for each patient were used to calculate the median fold change in expression of an RBP to account for extreme outliers. The results also emphasize that these high expression levels may be indicative of a major dysfunction of these RBPs in addition to dysregulation. For example, the mutated form of PHF6, which is implicated in various forms of leukemia, has higher expression. Alternatively, the change in expression may be a result of an upstream alteration in the regulatory mechanisms, for example NONO; another example is that NKRF expression is regulated by miR-301a [48]. The high expression of some of these RBPs may be the result of their normal physiological levels

being too low compared to a cancer context, as is the case for the proposed positive prognostic marker, RBM3. So, a natural question to ask is whether RBPs have some prognostic impact for cancer, starting from the trends that have been observed in this expression analysis.

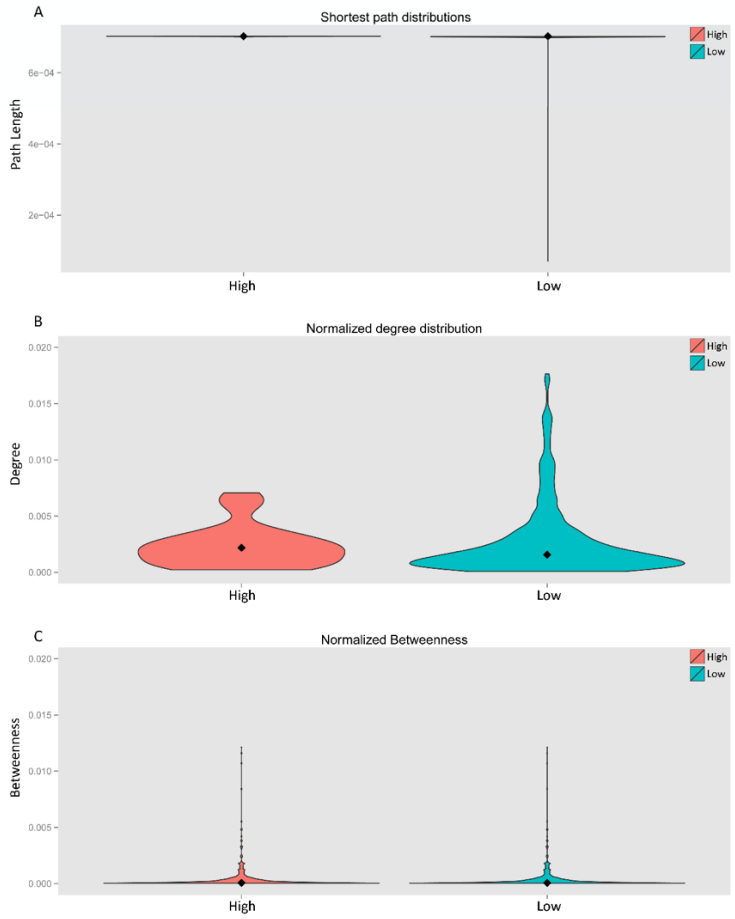


Figure 8: Comparison of normalized network metrics (closeness, betweenness and degree) between strongly upregulated (SUR) and non-strongly upregulated (non-SUR) RNA-binding proteins. The median values for each property are the same and there are no significant differences ($P > 0.05$, Wilcox test).

2.3.3 Strongly upregulated and non-strongly upregulated RNA-binding proteins exhibit significantly different within-group path lengths and variability in expression is related to the number of interactions

To identify further characteristics that differentiate SUR RBPs in cancer, I calculated the network properties of all the RBPs using a network constructed from the experimentally reported set of protein–protein interactions in the human genome obtained from the BioGRID database [25] (see Materials and methods). In particular, the shortest paths between pairs of proteins were computed within SUR and non-SUR RBP groups (that is, distances from SUR RBPs to SUR RBPs and distances from non-SUR RBPs to non-SUR RBPs) (Figure 9A). SUR RBPs were found to have significantly shorter path lengths to each other when compared to non-SUR RBP path lengths ($P < 2 \times 10^{-16}$, Wilcoxon test). Other network metrics such as normalized degree distribution, normalized closeness, normalized betweenness and mean path lengths for RBPs in each group were also calculated (see Materials and methods). However, there was no significant difference between SUR and non-SUR RBPs for these properties (Figure 8). This suggests that the interaction properties of an individual RBP (whether it is a hub and so on) do not relate to its dysregulation but rather the set of SUR RBPs are closely intertwined in the physical interaction network compared to the non-SUR RBPs. Although my observations on dysregulation are at the RNA level, it is possible to speculate, from the shorter path lengths observed, that the interaction network and crosstalk between SUR RBPs could also be perturbed in cancer genomes, with one or more of the SUR RBPs predominantly contributing to this perturbation.

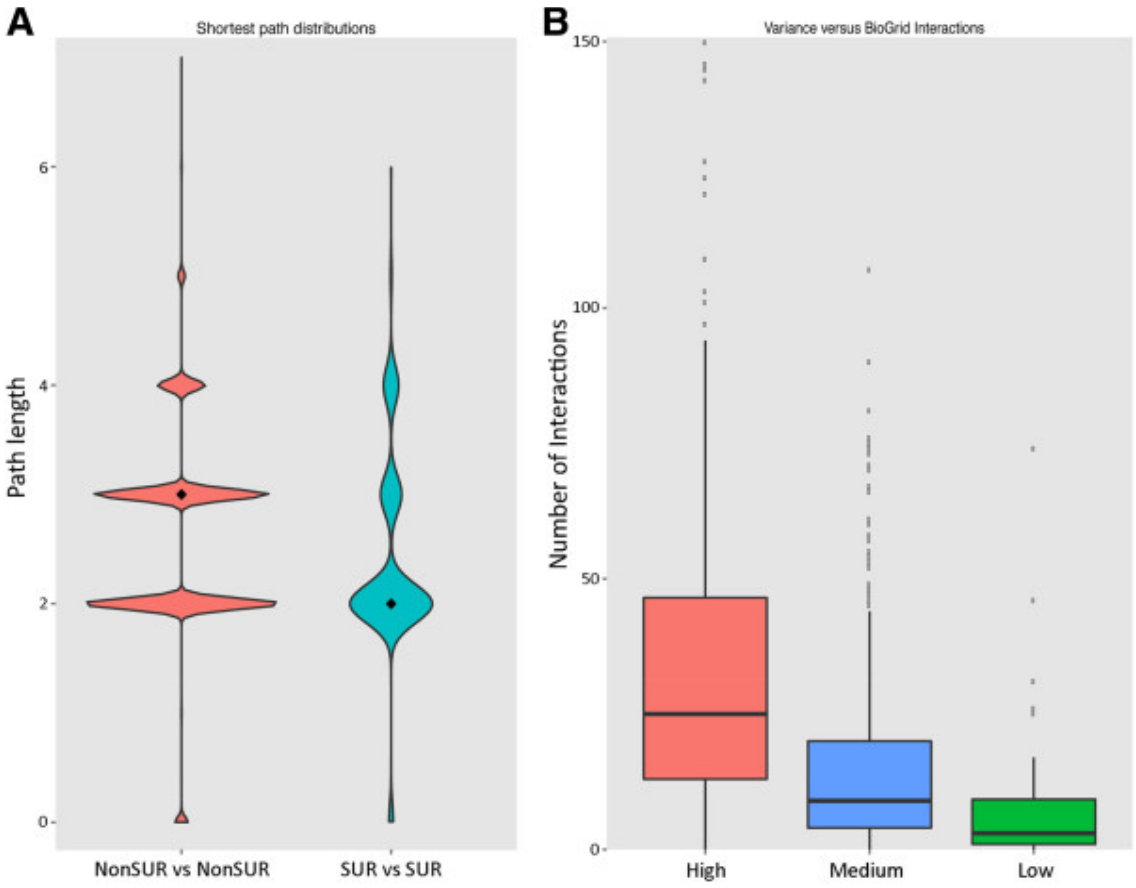


Figure 9: Interaction profiles of RBPs. (A) Distribution of shortest path lengths between every pair of RBPs belonging to SUR and non-SUR RBP groups using the protein–protein interactions documented in the BioGRID database [66], shown as violin plots. The width of each plot is the frequency distribution and the diamond is the median value for the category. SUR RBPs were found to have significantly shorter path lengths amongst themselves in comparison to non-SUR RBPs ($P < 2 \times 10^{-16}$, Wilcoxon test). **(B)** Box plot showing the number of interactions identified in BioGRID data for RBPs classified by variability levels defined by observed percentiles. The higher the variability for a RBP, the higher the observed number of protein interactions ($P = 9.247 \times 10^{-16}$, low vs medium; $P < 2.226 \times 10^{-16}$, low vs high; $P = 6.6556 \times 10^{-16}$, medium vs high, KS test).

Since the analysis of the shortest path lengths between RBPs from SUR and non-SUR groups suggested that the particular protein interaction partners of RBPs might play an important role in mediating or cascading the effect of dysregulation, I rationalized that the protein complex size and a RBP's occurrence frequency in protein complexes would be related to their sensitivity to dysregulation. RBPs long have been known to form protein complexes, and if a key component within a complex is dysregulated or malformed, it would affect its overall functionality. If a SUR RBP was very prolific one would expect that many patterns of dysregulation would occur downstream as a result of the formation of a faulty complex. Furthermore, if these SUR RBPs participate in smaller complexes, it may be that their dysfunction will not be regulated or counteracted by other members within the complex. From the CORUM data [27] (see Materials and methods), five SUR RBPs were identified and 172 non-SUR RBPs were identified. I found that for the two classifications of RBPs (SUR vs non-SUR), there were no significant differences in distributions for either complex size or complex frequency nor was there any correlation with expression levels. While the current coverage of the experimentally characterized human protein complexes is very limited, these results indicate that SUR and non-SUR RBPs do not have significant differences in terms of their protein complex membership.

I next asked whether the variability in expression levels of an RBP across cancer patients is different between SUR and non-SUR RBPs. To address this question, breast cancer was chosen as the disease model due to the fact that it is the cancer with the most patient samples in TCGA and would naturally be the most robust dataset for identifying variation in the fold changes in expression levels of a RBP. SUR and non-SUR RBPs did

not exhibit significantly different expression variation ($P = 0.1212$, KS test), which was measured as the median absolute deviation (MAD) in the expression fold changes between healthy and cancerous tissue across all the patients (see Materials and methods). However, an analysis to test the relation between expression variation and the number of protein interactions of an RBP revealed that the higher the expression variation, the higher the number of protein interaction partners of the RBP (Figure 9B). Indeed, a significant difference was noticed in the number of interactions in the classified levels of variability for RBPs ($P = 9.247 \times 10^{-16}$, low vs medium; $P < 2.226 \times 10^{-16}$, low vs high; $P = 6.6556 \times 10^{-16}$, medium vs high, KS test). In contrast, TFs did not exhibit such significant differences in the number of interactions with the classified levels of variability ($P = 0.8931$, low vs medium; $P = 0.0014$, low vs high; $P = 0.01$, medium vs high, KS test). However, for non-RBPs a significant difference was found between medium and high as well as between high and low levels of variability ($P = 0.7519$, low vs medium; $P < 2.2 \times 10^{-16}$, low vs high; $P < 2.2 \times 10^{-16}$, medium vs high, KS test). The observation that the higher the variability in expression of a RBP the more interactions it has, suggests that fluctuating RBPs whose expression is not tightly controlled might have more promiscuous (non-specific) protein interactions (and protein complexes) thereby leading to RNA off-targets at post-transcriptional level. My results also suggest that such dysregulation may be suppressed or is minimal due to the lower number of interactions for RBPs with less variability in expression. The analysis here has focused on the RNA expression levels of RBPs though it is likely that there will be influences from diverse post-transcriptional regulatory phenomena like alternative splicing, translation control and post-translational modifications, which will affect the ultimate protein levels. My

observations do provide evidence that RBPs with high variability in expression have a higher number of protein interactions.

2.3.4 Survival contributions of RNA-binding proteins in breast cancer is related to network proximity to strongly upregulated RBPs and variability in expression across patients

Based on the observation that SUR and non-SUR RBPs significantly differ in their within-group shortest path lengths, I questioned whether the path length of an RBP within the protein–protein interaction network might contribute to its prognostic impact for a cancer. Each RBP was ranked in each classification based on the mean path lengths to all connected nodes in the BioGRID protein interaction network and also computed the mean shortest paths to other nodes belonging to SUR RBPs and non-SUR RBPs. This allowed the construction of profiles for overall mean path lengths, lengths within-group for members of the SUR and non-SUR groups, and between the groups. The top five genes with the shortest and longest mean path lengths, and a randomly selected set of genes with intermediate mean path lengths, were selected for the survival analyses (Figure 10) (see Materials and methods). As the mean path lengths between SUR RBPs increased, their contribution to prognostic impact increased. This suggests that SUR RBPs with longer path lengths, that is, those with higher network distances with respect to other SUR RBPs, are more likely to contribute independently to survival as they might influence a larger fraction of the dysregulated network of SUR RBPs. On the other hand, when non-SUR RBPs were sorted by rank based on their mean path lengths with respect to SUR RBPs, I found the opposite trend. This suggests that non-SUR RBPs with shorter distances to SUR RBPs contribute to the perturbation of an important section of the RBP

protein interaction network. In particular, if a non-SUR RBP has a shorter path length, it has a good prognostic impact on survival for patients with breast cancer due to its lower expression. SUR RBPs are potentially in a malfunctioning state, and the closer a RBP is to them, the more the prognostic impact influenced by the SUR RBP interactions.

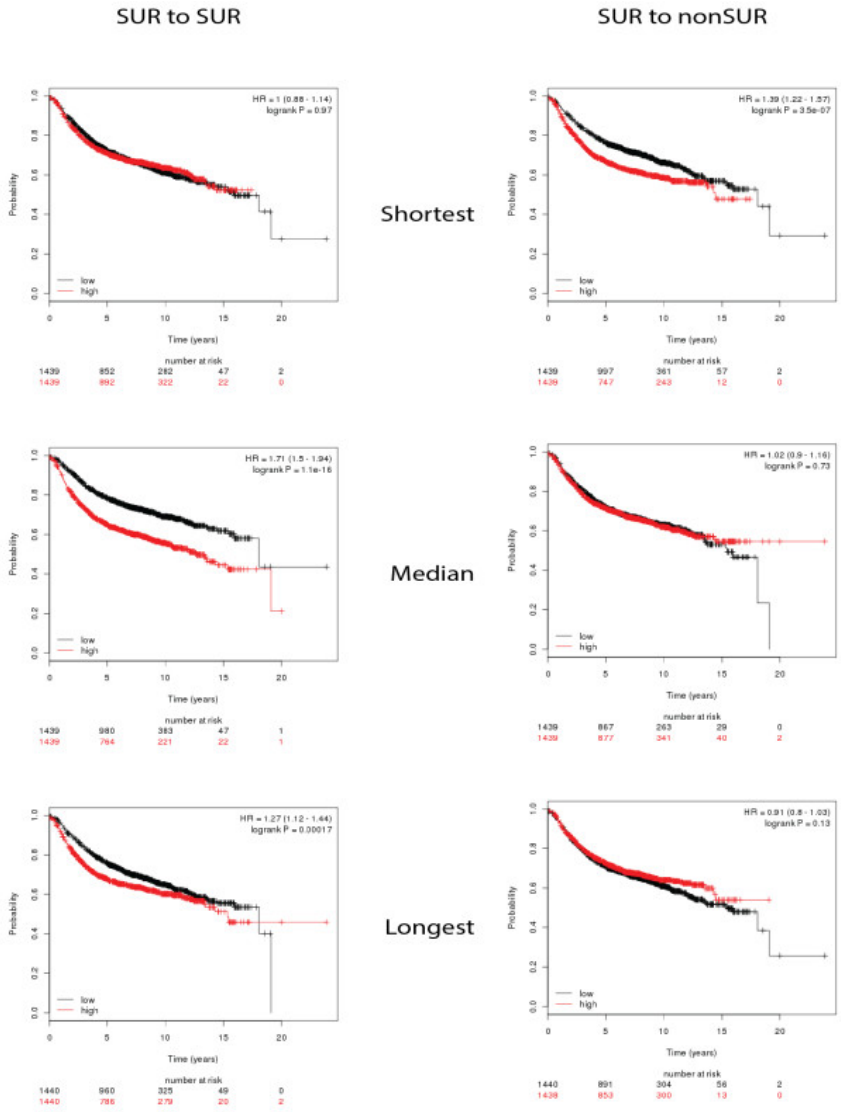


Figure 10: Survival of patients with breast cancer for different expression levels and path lengths for within and between expression groups of RNA-binding proteins. SUR (left) and non-SUR (right) survival for a sample of five RBPs classified by path length (shortest, median or longest). Curves in red are survival plots for patients with

enhanced expression of the selected genes based on more than 1,800 patients' expression profiles from the KM plot [28]. The within-group path ranking for SUR RBPs suggests that as the mean path lengths increase the contribution of the SUR RBPs in prognosis tends to increase. While between groups, RBPs having shorter path lengths to a SUR RPB contribute the most to prognosis.

I then compared the overall significance of the Kaplan–Meier *P* values (-log[*P*]) for groups of RBPs classified by their level of dysregulation (SUR versus non-SUR) and their levels of variability in expression across patients (high, medium and low variability determined by quartiles, see Materials and methods) in breast cancer (Figure 6). For both RBPs and non-RBPs, there was no significant difference between SUR and non-SUR genes in terms of prognosis for survival ($P = 0.12$ and $P = 0.06$, KS test) (Figure 11A,B). However, when I compared the significance of the *P* values for survival between SURs from RBP and non-RBP groups we found them to be significantly different ($P = 0.05$, KS test). In the comparison between variability levels of genes in RBPs, there was no significant difference between the Kaplan–Meier (KM) analysis significance levels ($P = 0.945$, low vs medium; $P = 0.3566$, low vs high; $P = 0.1478$, medium vs high, KS test) (Figure 11C). For non-RBPs, the levels of variability did have a very significant difference in the significance of KM-plotter survival *P* values ($P < 2.226 \times 10^{-16}$, low vs medium; $P < 2.226 \times 10^{-16}$, low vs high; $P = 6.6556 \times 10^{-16}$, medium vs high, KS test) suggesting that, in general, the higher the expression variation of a group of genes, the smaller is their contribution to prognosis for survival (Figure 11D). While there was no significant difference in RBPs I did observe a similar weak trend where the lower the variance in expression across patients, the greater the KM-plotter significance. A highly

variable RBP has less effect on survival because it could potentially be regulated by a number of other factors and could be the result of an indirect effect, whereas low variability RBPs have a less but more direct effect on the prognosis for an individual and hence could be the actual drivers. This also corroborates our notion after observing variability versus the number of protein interactions (Figure 7B). More generally, my results suggest that while I observe a larger proportion of SUR RBPs, their elevated expression alone does not necessarily mean they have a direct effect on positive or negative prognoses.

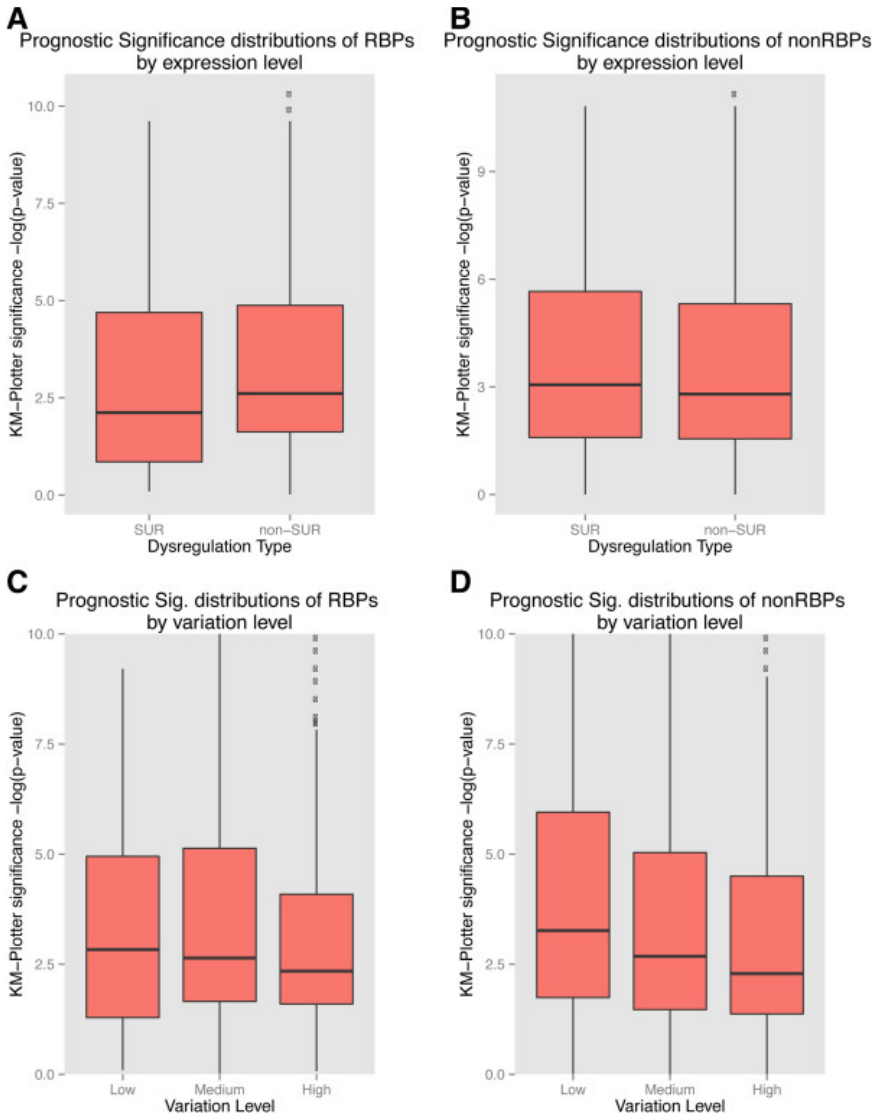


Figure 11: Comparison and distribution of prognostic impact based on expression dysregulation and expression variability in breast tissue. RNA-binding proteins **(A, C)** and non-RNA-binding proteins **(B, D)** were categorized based on their level of dysregulation as healthy or cancer expression (SUR or non-SUR) and the variability of expression levels (high, medium or low) in patients with breast cancer. The statistical significances for the differences in the distributions of prognostic impact are discussed in the main text. KM, Kaplan–Meier; RBP, RNA-binding protein; Sig., significance; SUR, strongly upregulated.

2.4 Conclusions

In this study, I investigated the gene expression profiles of RBPs in healthy humans for 16 tissues and found that RBPs are consistently and significantly highly expressed compared to other classes of genes (non-RBPs) as well as in comparison to well-documented groups of regulatory factors like transcription factors, miRNAs and lncRNAs. This, in concordance with previous research, emphasizes their importance in post-transcriptional regulatory control across all the tissues. To understand the expression profile changes in a disease state for hundreds of RBPs in the human genome, I obtained analogous RNA-sequencing-based expression data for a total of 2,876 patient samples spanning nine cancers from TCGA and calculated a log-ratio for expression between cancer and healthy states. I showed that there is a unique signature of approximately 30 RBPs that had significantly increased expression levels across six out of nine (two-thirds) cancers profiled. These could be clearly labeled as a set of SUR RBPs delineating them from the rest of the RBPs based on the change in expression levels. This proportion of SUR RBPs in the RBP population is greater than the proportion of SUR non-RBPs suggesting for the first time that the expression levels of a significant fraction of the RBPs are affected in cancerous states. Analysis of the protein–protein interaction network properties for SUR and non-SUR group of RBPs, suggested that the shortest path length distributions between SUR RBPs is significantly lower than that observed for non-SUR RBPs. This observation together with survival analysis based on path lengths suggests that not all the SUR RBPs might be directly implicated in cancer but rather that a cause-and-effect relation might hold between some of the SUR RBPs. This observation was further supported by the fact that the higher the expression variation of a RBP in breast

cancer patients, the higher the number of protein–protein interactions. This indicates that fluctuating RBPs whose expression is not tightly controlled (with differing fold changes in expression levels across patients) might be involved in more promiscuous (non-specific) protein interactions thereby leading to variable RNA off-targets at the post-transcriptional level.

To further determine the prognostic impact in breast cancer patients the SUR and non-SUR RBPs were ranked based on path length. The two RBP groups had different distributions. As the mean path lengths between SUR RBPs increased their contribution to prognostic impact increased, suggesting that SUR RBPs with higher network distances with respect to other SUR RBPs, are more likely to contribute independently to survival as they might influence a larger fraction of the dysregulated network of SUR RBPs. In contrast, when a non-SUR RBP had a shorter path to a SUR RBP, there was a significant prognostic impact. This suggests that they are closer to the actual contributors of pathogenesis at the post-transcriptional level; however, the longer the path lengths, the weaker the prognosis. To gain further insight into the contribution of these subsets of RBPs in the development of and survival with cancer, I compared the overall significance of the Kaplan–Meier *P* values (-log[*P*]) for groups of RBPs classified by their level of dysregulation (SUR vs non-SUR). This analysis revealed no significant differences between groups of SUR and non-SUR RBPs in terms of their prognosis for survival. However, generally, the higher the expression variation across patients, the lower the prognostic impact of the protein. The results suggest that RBPs from the signature set with lower variation in expression levels across patients might be good starting points for studying the effect of RBPs in cancer pathogenesis since SUR RBPs with large

expression fold changes might be downstream or there might be indirect effects. Additionally, common factors that are dysfunctional along the shortest paths in the protein interaction networks of SUR RBPs could also provide clues for potential drug targets as they can act as regulators for rewiring the post-translational landscape of RBPs thereby affecting RNP complex formation. With increasing efforts to uncover the binding sites of RBPs in higher eukaryotes using a variety of high-throughput approaches [73,74], it should also become possible in the near future to study the differences in the target RNA pools between healthy and cancer genomes for several of these SUR RBPs. This would provide a global picture of the affected post-transcriptional regulatory networks. The global integration of networks governed by post-transcriptional players like miRNAs and RBPs together with signaling networks can provide a comprehensive picture of the cause of the dysregulation in these RBPs, which can be used to tease apart the contributions of local malfunctions and those due to an upstream or downstream effect in the cellular networks.

3. Bottom-up, integrated -omics analysis identifies broadly dosage-sensitive genes in breast cancer samples from TCGA

3.1 Background

The scientific literature is replete with papers highlighting the complex interplay between chromosomal instability, aneuploidy, and cancer (e.g. [75] [76] [77] [78]). Aneuploidy, the state of having other than the canonical or "euploid" number of chromosomes - for humans, 46 - is with only rare exceptions (Downs syndrome, Trisomy 18) lethal in human embryonic development [79]. By contrast, aneuploidy is observed with very high frequency in cancer, leading the eminent German biologist Theodor Boveri to speculate as early as 1902 [80] that aneuploidy might have a causative role in the disease.

Despite previous investigations, there are still important questions. Is aneuploidy a cause or a side-effect of cancer? If the former, what factors associated with aneuploidy contribute to cancer cell fitness? Are there deleterious impacts of aneuploidy in cancer and how are they mitigated during tumorigenesis? More generally, what is the broader impact of aneuploidy on gene expression and resulting phenotypes?

DNA studies have found that amplification of genomic arms such as 20q and 8q [81] [82] occur with high prevalence and have been correlative with cancer severity. Understanding how these amplifications impact changes in gene expression and protein production is of great interest. The conventional wisdom regarding gene transcription and translation has been that "dosage" *generally* correlates with product: DNA to RNA

to protein. Indeed, a recent report finds no evidence for widespread dosage compensation in yeast [83].

It is customary to use mRNA transcript abundance to identify disease-associated genes, but the impact of mRNA abundance on protein production is poorly understood. Correlational methods yield weak associations, even when considering protein half-lives and other chemical properties [84–87]. Other efforts have been made to integrate mRNA dynamics (half-life and fold energy) and RNA Binding Protein (RBP) interactions with expression data in *S. cerevisiae* and *S. pombe* to aid in predicting protein production from gene expression. [88] Illustrates how sequence elements (sequence lengths, secondary structures, etc.) were used to identify protein abundance variations. Understanding how DNA, RNA, and protein interact is a non-trivial task, but considering any of these features in isolation may yield sub-optimal results. This understanding could provide crucial details about tumorigenesis, cancer evolution, and may hold clues to potential cancer treatments.

In 2015 approximately 40,000 women died of breast cancer in the US alone [89]. In an effort to better profile and understand cancer, large public efforts have been initiated to gather patient data and comprehensively investigate it. The Cancer Genome Atlas (TCGA) collects data for patients across 34 types of cancer profiled using a wide array of '-omics' platforms [90]. The unprecedented availability of cancer data, like TCGA, affords insights into the genomic foundation of these lethal diseases.

Here big data methods were applied in a systematic fashion to observe the impact of DNA dosage on mRNA transcript levels and subsequent protein concentrations. I identify the prevalence of dosage compensation in TCGA breast cancer samples (BRCA), highlight dosage-sensitive genes, and investigate the role of these genes in cancer cell line survival.

3.2 Material and methods

The data used in this study has been downloaded from multiple resources, including TCGA [19], Clinical Proteomic Tumor Analysis Consortium (CPTAC) [91], the Catalogue of Somatic Mutations in Cancer (COSMIC) [92], and Achilles short hairpin RNA or small hairpin RNA (shRNA) [93]. The data and processing approaches are briefly described below. Figure 12 illustrates the overall workflow.

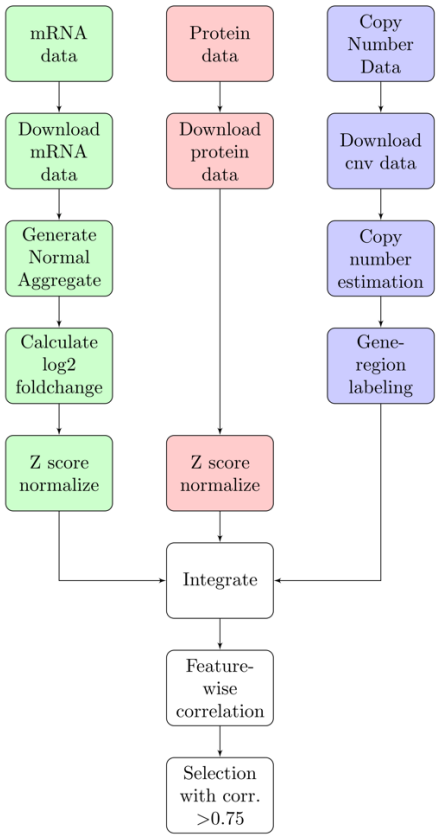


Figure 12: Bottom-up, integrated analysis workflow. Visual representation of the analytical workflow for identifying broadly dosage-sensitive genes. Green portions represent mRNA-based steps, red protein, blue CNV. Integrated and filtering steps are white. Briefly, data were acquired from their sources, joined with metadata, normalized, integrated, then filtered.

3.2.1 The Cancer Genome Atlas (TCGA)

RNAseq V2 data of 114 normal control patients and 1102 patients with breast invasive carcinoma (BRCA) were downloaded from TCGA. For each of the 20532 genes of each patient, the median of 114 normal values was used as an estimated baseline, which is noted as the *Normal median RSEM*. The fold change of each gene, $\Delta_{exp}(gene)$, was calculated for every patient as:

$$\Delta_{exp}(gene) = log_2 \frac{Cancer\ RSEM\ (gene)}{Normal\ median\ RSEM(gene)} \quad \text{(eq. 1)}$$

The corresponding patient metadata was downloaded and mapped based on sample IDs extracted from the TCGA barcode. Level 3 TCGA copy number variant (CNV) data was extracted for all available patients. The TCGA CNV pipeline transforms a CNV value into a segment mean, where $Segment\ mean = log_2(CNV/2)$. A copy number can be derived from the segment means by calculating $2 * (2^{segment\ mean})$. With this, the diploid regions will have a segment mean value of zero, amplified regions will have positive values, and deletions will have negative values.

3.2.2 Clinical Proteomic Tumor Analysis Consortium (CPTAC)

Mass spectrometry (MS) data for breast invasive carcinoma was downloaded from CPTAC; these abundances were reported as the $\log_2$-ratio of the expression of the sample to a common, healthy pool [94]. Patient mRNA, protein, and CNV data was matched using TCGA barcode. Gene identifications for MS data were made using previously established methods [95] and were provided by CPTAC. mRNA, protein, and CNV data for a given gene were joined by gene symbol. The unshared relative protein abundance was matched to 106 patients with mRNA and protein abundance data [91].

3.2.3 Gene amplification and deletion

Segment regions were mapped to the UCSC genome coordinates for the hg38 build of the human genome. For regions that covered multiple genes, the segment means counted for each gene. For genes with multiple calls, the maximum value was kept.

For genes across all 106 samples with protein and mRNA data, there were 1,052,345 segment means in total. The average segment mean was 0.12 ± 0.02. A gene is

defined as amplified if its segment mean is greater than 0.2 and deleted if it is less than -0.2 [96]. Doing so there are 837,531 normal segments in the patients, 213,361 segments are amplified, and 1,453 have deletion events. Of those, 9835 genes were uniquely identified as normal, 9831 as amplified, and 1247 as deleted. Interestingly, only 15 patients of the 106 had no deletions.

To normalize protein and mRNA expression to a similar scale, a z-score normalization for protein and mRNA fold changes was performed as follows:

$$z = \frac{x_{\Delta_{exp}} - \mu_{\Delta_{exp}}}{\delta_{\Delta_{exp}}} \quad \text{(eq.2)}$$

Where $x_{\Delta_{exp}}$ represents the expression of a given gene, $\mu_{\Delta_{exp}}$ the mean expression for the data set, and $\delta_{\Delta_{exp}}$ the standard deviation.

3.2.4 Cancer gene profiling

The Cancer Gene Census was downloaded from COSMIC [92]. Genes labelled as amplified were selected and mapped to the results of the integrated genomics analysis to annotate correlational signatures. The Pearson correlation coefficient was calculated between protein abundance fold-change, mRNA fold-change, and CNV amplification. Any gene with all correlational scores above 0.70 is called a "Broadly Dosage-Sensitive Gene," or **BDSG**. The stringent cutoff was selected to emphasize genes as very unique. Generally, there was poor correlational concordance among the features (Figure 13). This is not to say the genes below this threshold may not be informative, but they are not exemplars of this particular genomic conservation.

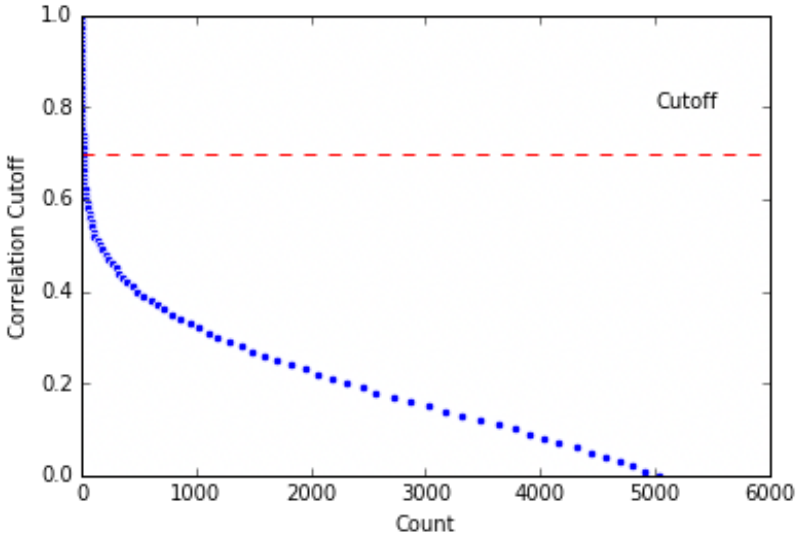


Figure 13: Correlation coefficient cutoff. Scatter plot of correlation threshold cut offs and subsequent member counts. The red, dotted line indicates cutoff used for BDSG identification for the study.

3.2.5 Achilles shRNA

Achilles shRNA knockdown data was downloaded and subset for the BDSG genes, a selection of housekeeping genes [97] and genes acting as oncosuppressors or oncogenes. shRNA hairpins for each gene were selected based on second-lowest log-ratio to avoid false positives. Hierarchical clustering was performed via Python clustermap function on the genes, as well as the cell types. Additional hierarchical clustering was performed excluding any cell types not related to breast models.

3.3 Results and discussion

3.3.1 Genomics and protein analysis

RSEM distributions for the 106 breast cancer patients for 20531 genes were plotted for both the cancer and healthy samples. Initial observations of RSEM values show similar quantiles suggesting that global expression distributions are similar between

tumor-matched normal and tumor samples (Figure 14). This alleviates concern for the impact of any batch-effect or other temporal anomalies in sample processing.

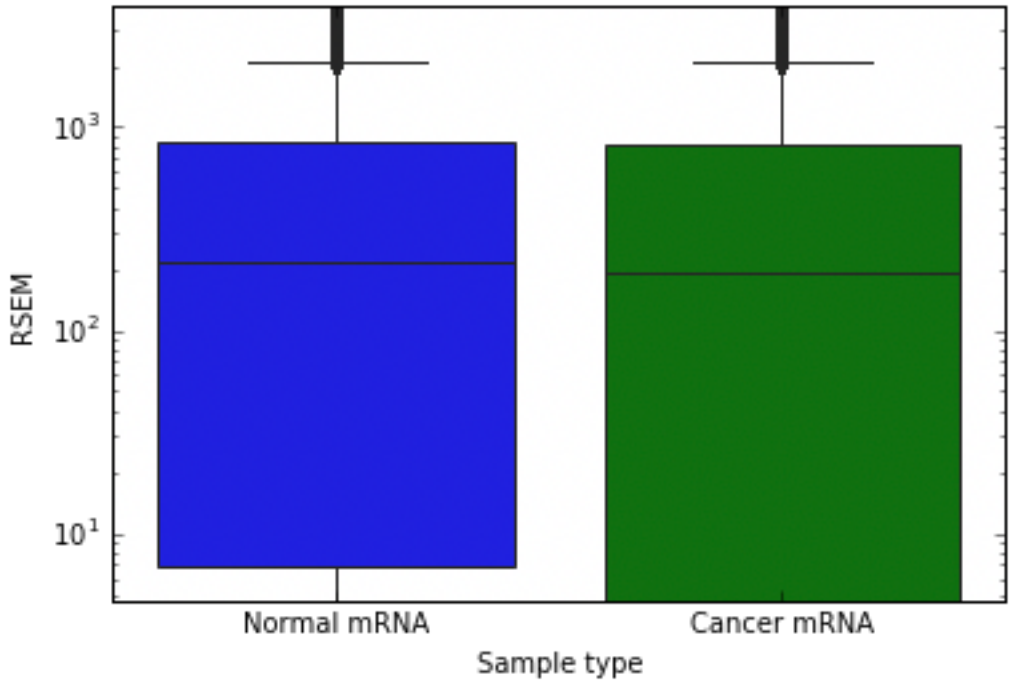


Figure 14: Tumor-matched normal and tumor sample expression distributions. mRNA expression distributions for tumor-matched normal and tumor samples. Distributions are similar suggesting no batch effect or temporal confounder. This does not illustrate changes of a singluar genes expression between the two states, rather summary information of expression.

Wilcoxon signed-rank test of gene expression between healthy and cancer samples detected that over 10,000 genes had significantly different expression ($p < 0.00005$). This emphasizes that the cancer transcriptome varies dramatically from normal tissue.

The $\log_2$ fold-change of mRNA from cancer to healthy was calculated as (eq.1), described with a mean mRNA fold change of $-9.0 \times 10^{-2} \pm 1.5$ (max = 17, min= – 16).

This indicates a class of genes that have minimal change and another subset that shows large changes in expression as indicated by the distribution (Figure 15).

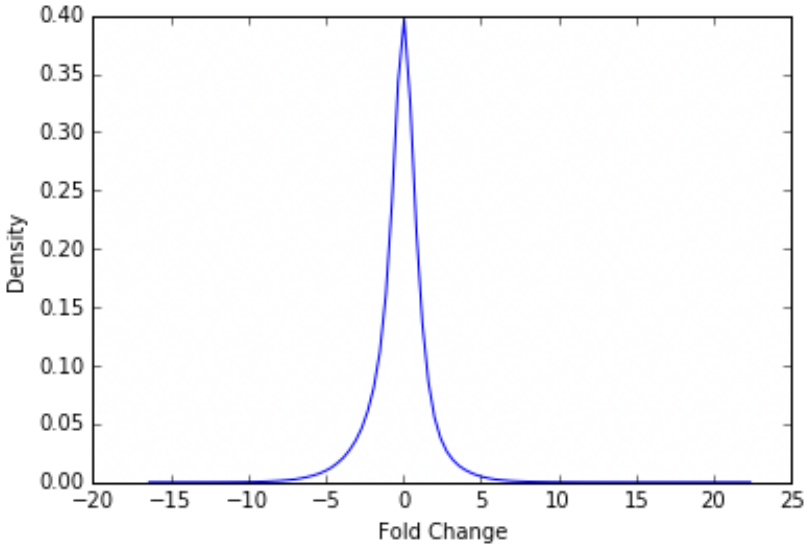


Figure 15: mRNA log2 fold change from normal median to cancer. mRNA log2 fold change distributions illustrate classes of genes within patients that have high degrees of alteration. In this instance we can now identify genes with large log2 fold change differences from cancer to healthy states.

Fold changes were plotted for protein expression. For proteins, the mean fold change was $3 \times 10^{-2} \pm 0.67$ (max = 8.5, min = – 6.5). Since the dynamic range of these effects are not equivalent in both datasets, they were z-score normalized. After normalization, the overall distributions of mRNA and protein fold changes were not statistically different ($P = 0.96$, Wilcoxon Test). Therefore, my normalization practice is sufficient to integrate the datasets and attempt to find relationships. These relationships are visualized in Figure 16. The strikingly poor correlation between the two features emphasizes the difficulty in accurately inferring protein levels from given mRNA expression values. A D'Agostino's K-squared test of the mRNA and protein correlations led to the rejection of the null hypothesis that the data was Guassian ($P <$

0.0005). Further examination of QQ plots conveys a trend for a mariginally lighter left tail and a heavier right tail to the distribution.

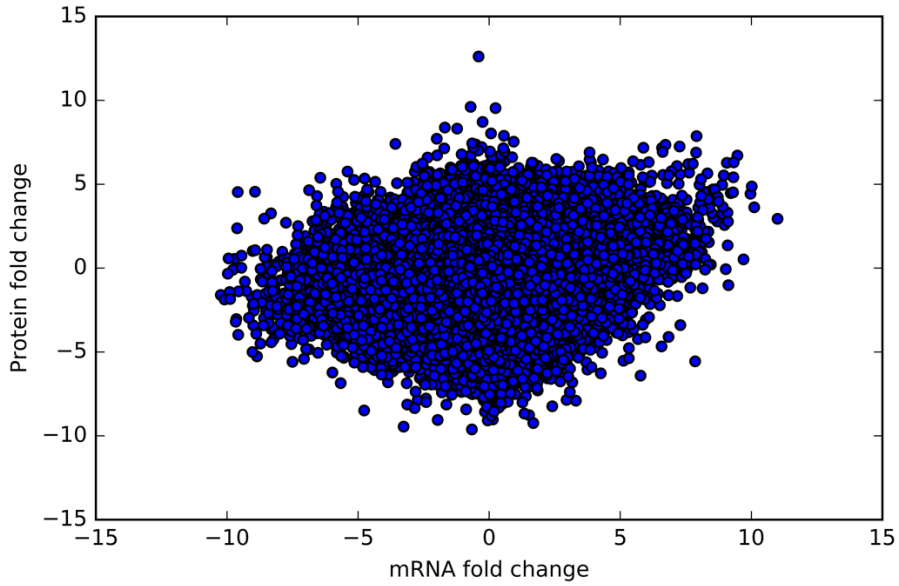


Fig 16: mRNA fold change versus protein fold change. This represents the matched protein (y-axis) and mRNA (x-axis) z-score normalized, $\log_2$ fold changes of disease to healthy samples for 106 patients and 20531 genes. Poor correlation amongst the points suggests that mRNA changes do not necessarily predict protein production changes.

3.3.2 mRNA and protein dysregulation relative to copy number variation

There is a strong positive trend for correlation of mRNA fold change with segment means. This is intuitive, given if there is an amplification or deletion of a genomic region transcriptional activity is impacted.

A surprising result is the volume of anti-correlated genes when comparing mRNA versus protein fold-changes. There are 185 genes out of the 9835 with weak, negative correlations between mRNA and protein fold-changes. A negative correlation suggests that a regulatory mechanism is either additionally suppressing or enhancing transcript

abundance. There are a number of genes which have negative correlation. These are of relatively low magnitude and lack statistical significance, but there may be underlying biological mechanisms at play that may be of interest in future studies. This notion of regulatory interactions is emphasized again by observing the correlation of protein changes to segment means (Figure 17). In this case there are generally weak correlations, supporting the concept that precursor genomic features are often not a reliable predictor of protein concentration [98].

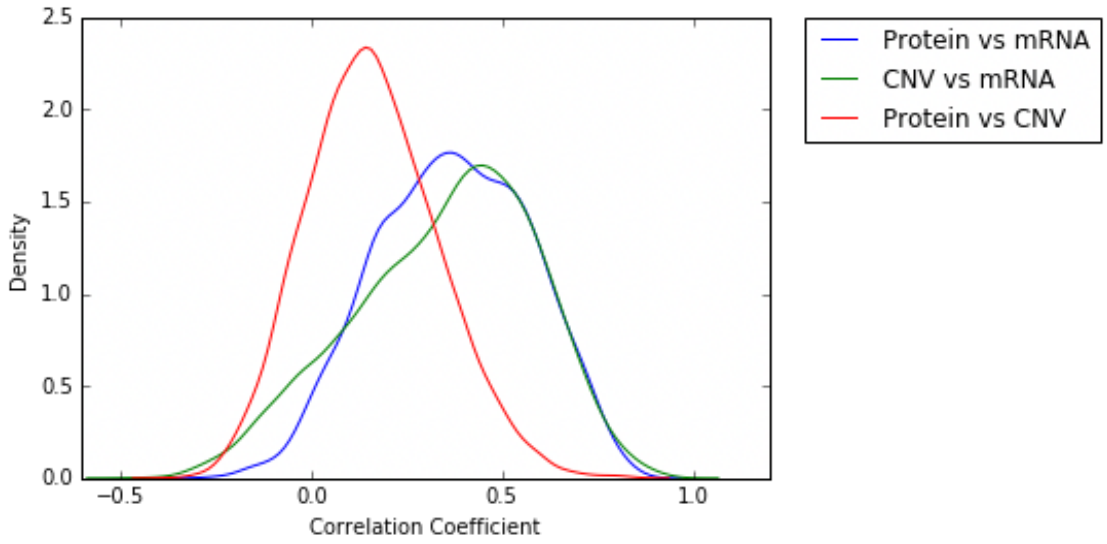


Figure 17: Correlation coefficient distribution. Overall correlation coefficient distributions between genomic features emphasizes the decoupling of mRNA and CNV from overall protein production in the TCGA breast cancer samples.

Selecting genes and subsequent data based on the 20q chromosomal arm – a locus known to be frequently-amplified in breast cancer [99] exemplifies how DNA amplification, mRNA and protein abundance may be discordant. In the case of the gene SLPI for a sample (TCGA barcode: D8-A13Y), there is strong DNA amplification,

however mRNA and protein abundance is very low. SLPI encodes an antibody-producing transcript which antagonizes paclitaxel in ovarian cancer cells [100].

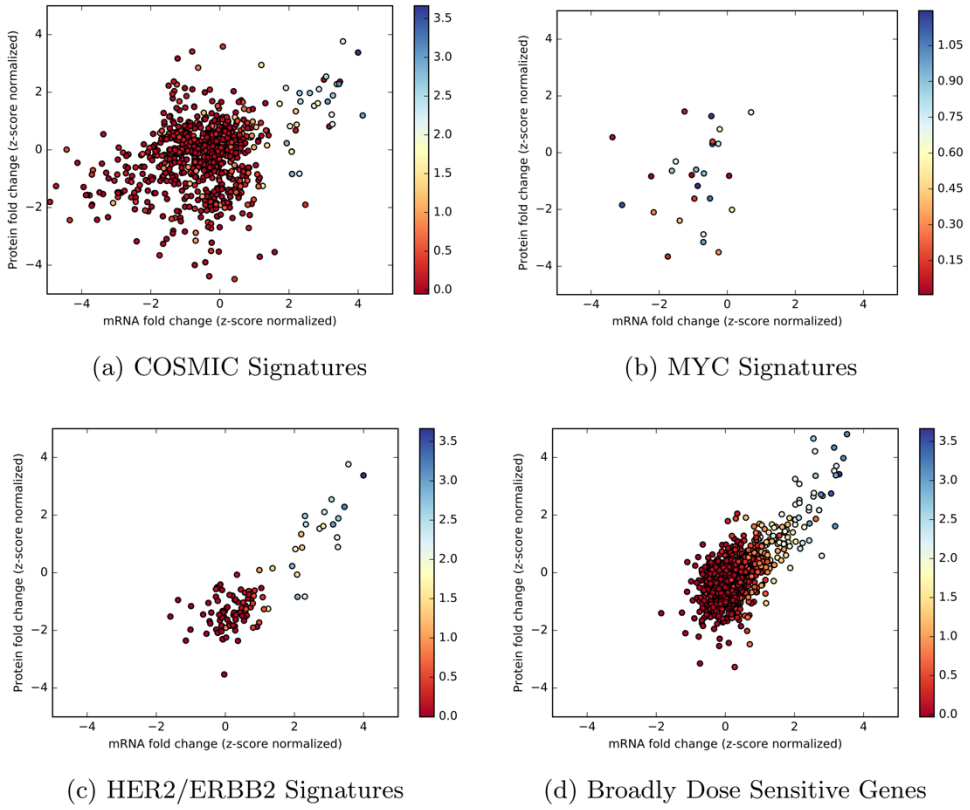
Conversely, there are cases where there is concordance between amplification, mRNA abundance, and protein abundance. In the case of RIMS4, there are very high fold changes and amplification in certain members of the cohort. KM-survivability analysis [28] based on microarray data of over 1000 breast cancer patients indicates that high RIMS4 expression has a positive prognostic impact ($P < 4.0 \times 10^{-7}$).

3.3.3 Cancer gene profiling identifies broadly dosage-sensitive genes (BDSGs)

The mRNA, protein, and CNV data for genes labeled as amplified in breast cancer from COSMIC is in Figure 18A. Genes from across the genome that meet the Pearson correlation criteria (i.e., all correlations above 0.70) are displayed in Figure 18D and listed in Table 2. Among the genes in Table 2, ERBB2 (HER2, Figure 18C) is a member of this group and is a well-known oncogene. In both Figure 18C and Figure 18D, there is a strong dosage sensitivity that is atypical across the genome. The remaining 11 genes in Figure 18D are not identified in COSMIC at all. GRB7 is a growth factor receptor that overlaps with HER2 pathways, and coexpresses with it in esophageal cancer [101]. RPS6KB1 is a kinase whose alterations have been associated with an increased risk of colorectal cancer [102]. These broadly dosage-sensitive genes (BDSGs) are observations on a seemingly rare conservation of the central dogma, yet they have minimal functional annotations in the scientific literature.

Symbol	Location
ERBB2	17q11.2-q12
GBAS	7p12
GRB7	17q12
HEATR6	17q23.2
LANCL2	7q31.1-q31.33
PDSS2	6q21
PPFIA1	11q13.3
PPME1	11q13.4
RPS6KB1	17q23.1
SUMF2	7q11.1
TACO1	17q23.3
UBE2Z	17q21.32

Table 2: Broadly Dosage-Sensitive Genes (BDSGs)



(a) COSMIC Signatures

(b) MYC Signatures

(c) HER2/ERBB2 Signatures

(d) Broadly Dose Sensitive Genes

Figure 18: Protein vs mRNA fold changes with CNV amplification. Figures represent the z-score normalized, log2 fold change of mRNA (x-axis) versus protein (y-axis) and are colored by CNV segment mean values for samples from patients in the TCGA Breast Cancer dataset selected for: (a) Genes in the COSMIC database labelled as amplified, (b) known oncogene MYC, (c) known oncogene and BDSG HER2/ERBB2, and (d) all BDSGs. The trend in Figure 18A exemplifies how oncogenecity does not always correlate with dosage-sensitivity.

3.3.4 shRNA data defines the role of BDSGs in cancer cell line growth

Figure 19 shows results of testing the impact of BDSGs on cancer cell line growth using shRNA. TUBB is a common housekeeping gene and the signature illustrated in 19A is typical of this role. When knocked down by shRNA there is a very deleterious

effect on cancer cell line growth and it stands out as a singleton in both heatmaps. In Figure 19A PPME1 and UBE2Z consistently behave as tumor suppressor genes (TSGs) across all cell types; their silencing promotes cell viability. In breast cancer cell lines these genes cluster closely with PTEN and RB1 which were included as typical breast cancer TSGs. In contrast, GRB7 and RPS6KB1 have a generally negative impact on cell line viability in Figure 19A. However, when considering just breast cancer-specific cell lines, these genes cluster closely and exclusively with ERBB2. The differences in clustering behavior suggest that, unlike PPME1 and UBE2Z, GRB7 and RPS6KB1 act as oncogenes very similar to ERBB2. In fact, GRB7 is co-located with ERBB2 and may be upregulated as an adaptation to HER2 (29). According to the breast-specific cell type clustering of shRNA data, BDSGs do not display a subtype-specific role. They are generally tumor suppressors or oncogenes across all breast cell lines.

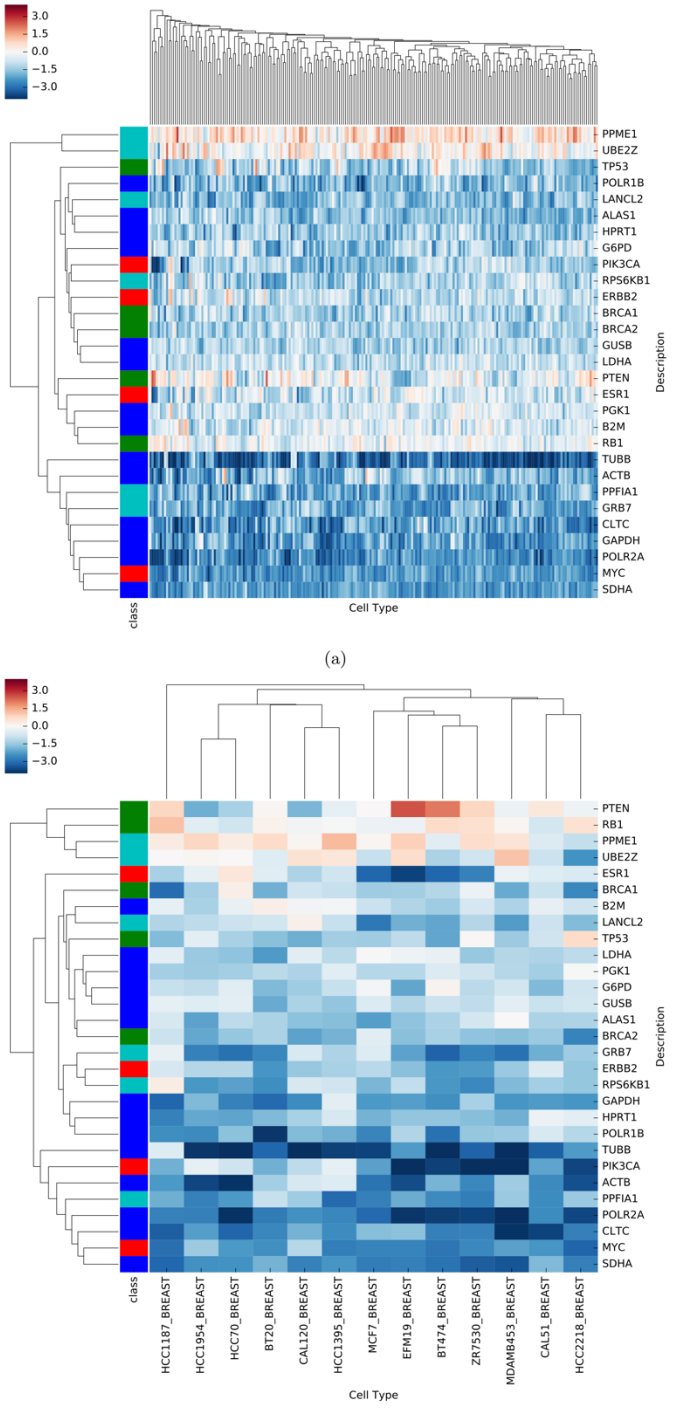


Figure 19: Heatmap and hierarchical clustering of shRNA knockdown. Section (a) represents all cell lines available in the Achilles project and (b) breast-specific cell lines. Rows are the selected genes: blue are housekeeping genes, red are oncogenes, green

oncosupressors, and light blue are BDSGs; columns are the cell lines. Red cell values represent cellular proliferation, blue cellular death, and white no change. The clustering of ERBB2 and two BDSGs (GRB7, RPS6KB1) in (b) suggests an oncogenic role in breast cancer. PPME1 and UBE2Z signatures in (a) and (b) suggest an overall oncosuppressive role.

3.4 Conclusions

In this section I have shown that in the TCGA breast cancer cohort there is widespread dosage compensation for the extensive aneuploidy that is observed. The dosage of DNA does not generally correlate well with mRNA, nor does the latter correlate well with protein levels. A total of 11 genes show strong correlation across all features (DNA/mRNA/protein); analogous to that of a well-known oncogene HER2 (ERBB2). These genes are referred to as "Broadly Dosage-Sensitive Genes" or BDSGs. It must be noted that they are much less characterized in the literature as to their role, if any, in cancer. I advocate further study of BDSGs to better understand their potential effects on cancer. This may lead to new therapies for cancer or biomarkers for improved cancer detection.

From shRNA data, I show that knockdown of these genes has an impact on cancer cell growth. I speculate that tumor cells adapt unusual ploidies to take advantage of amplifications and deletions that functionally implicate only subsets of genes. These tumor cells may compensate for the dosage of a large number of "passenger" genes. This may be a vulnerability that could be used for cancer therapy, for example by de-repressing mRNA and/or protein production from these passenger genes. This may leave

the tumor cell with potentially catastrophic levels of unneeded molecules or disrupted biological pathways.

I also caution that there may be significant pitfalls in drawing conclusions from a single type of genomics data. For example, gene expression (mRNA) data is widely-used to infer biological pathway activation, but Figure 16 suggests this would be extremely misleading for exploring protein levels of Cancer Gene Census genes in TCGA Breast samples.

4. Deep Learning and transcriptional signatures for identifying key differentiating genes in cancer histologies.

4.1 Background

In 2019, the American Cancer Society predicted that in the US alone, approximately 23,000 individuals would be diagnosed with ovarian cancer [104], with over half that number dying from the disease; and nearly an order of magnitude more patients being diagnosed with breast cancer. By building methodologies to detect how these cancer types behave differently at the genetic level, we can better understand the pathology of the disease as well as markers to uniquely identify it earlier.

The transcriptional landscapes of cancers are a complex series of interactions that are difficult to understand. As stated previously, it is difficult to determine if the changes occur as a result of having cancer or they are the driving characteristic of cancer. With developments in genetic engineering and screening, it becomes even more essential to understand and detail these subtle differences.

The efforts of the Clinical Proteomics Tumor Analysis Consortium allow an even more comprehensive view of the samples provided by the TCGA. Most importantly, the investigation can include the transition of genetic components from DNA to protein across an additional cancer type. We can also investigate how these multi-omic signatures can identify differentiating and identifying features.

The pervasiveness of deep learning methods and frameworks offer the opportunity to perform these explorations. Deep learning is being used extensively in the clinical diagnostic space [105–107]. These approaches provide the ability to analyze the complex relationship of features in dense datasets such as slide microscopy, medical

health records, time series data, or even sequencing data. Deep learning aids in the analysis of complex, multi-modal data. The available frameworks facilitate reproducible and transferrable results. Traditional machine learning and clustering methods struggle with correlative datasets, or those with large input parameters. Deep learning is more tolerant of these attributes and can afford quicker paths of analysis. It allows models that are composed of multiple processing layers to learn representations of data with many levels of abstraction [108]. Of interest to us is the ability to include the copy number, mRNA, and protein changes that occur from a normal to a disease state.

I propose examining the multiple facets of genomic data in breast adenocarcinoma and ovarian cancer to better understand the transcriptional activity in the disease states. I will compare genes whose gene expressions across multiple features (CNV, mRNA, protein) highly correlate in both subsets to determine if there are uniquely identifying signatures. Lastly, I will leverage deep models to classify the samples and utilize information and game theory approaches to determine feature importance and compare it across gene types.

4.2 Methods

4.2.1 Data selection

Breast adenocarcinoma mRNA RSEM read counts, copy number variation data, and associated metadata for 1097 tumor samples and 121 matched-normal samples from The Cancer Genome Atlas [19] were collected along with 375 tumors and six matched-normal samples from ovarian cancer. Similar data was collected for ovarian carcinoma tumor and matched-normal samples. From the Clinical Proteomic Tumor Analysis

Consortium, iTRAQ uniquely matched spectra were selected for overlapping TCGA BRCA and OV samples[94].

4.2.2 Data preparation

Expression, protein, and copy number variation data were prepared based on previous methods [109]. In brief, mRNA RSEM values were calculated as log2 fold changes using median, matched-normal sample expression as a healthy state. Copy number segment means were changed to copy numbers. mRNA and protein fold changes were z-score normalize to center means for further calculations.

The Spearman correlation was calculated across all samples between genomic datasets: copy number versus mRNA fold changes, mRNA fold changes versus protein fold changes, and copy number and protein fold changes. The higher correlation scores across data types will be used to identify broadly dosage sensitive genes (BDSGs) [109] and features to be used as input for model generation. Kolmogorov-Smirnov (KS) [110]tests were performed to compare distributions of correlational scores between histological data types.

Arithmetic means and standard deviations were calculated for the various fold changes and scores for exploratory observations. To assess any differences in expression changes Mann-Whitney U tests were performed between BRCA and OV datasets (for all genes, BDSGs as a category, and BDSGs per histology).

4.2.3 Machine learning models

A feed forward deep neural network was engineered using the Keras API to TensorFlow [111] to classify samples as belonging to one of two histologies: BRCA or OV. The model consists of three input channels, each with four dense layers,

concatenated into inputs to a final series of dense layers to predict the cancer type (Figure 20). For each layer besides the output layer, the ReLu activation function was used. Since the model is attempting a binary classification (either/or sample classification) the sigmoid activation function was used on the final, output layer. During training the ADAM function was used for optimization with a binary cross entropy loss function.

The input into the model are the mRNA and protein fold changes as well as the DNA copy numbers for 161 samples. This dataset was split into training and testing sets for model training and evaluation. Input genes were selected to train three models: all genes, BDSG, and randomly selected non-BDSG. For the all gene model, there were 10,548 genes; 74 for the BDSG-based model; and the random model was selected from 74 random, non-BDSGs.

For comparison to traditional machine learning methodologies a random forest model with 1000 estimators was generated using the Python SciKit Learn Random Forest module [112]. Classification results were generated from the same input sets as the deep learning model.

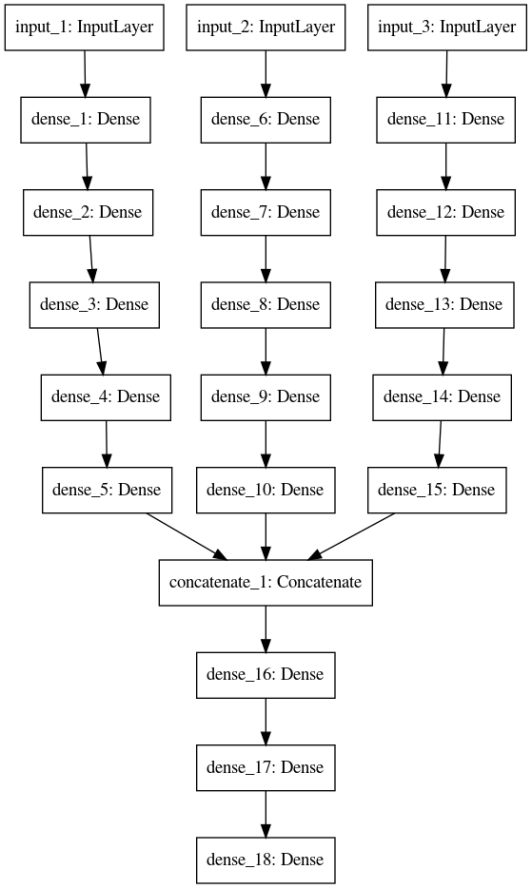


Figure 20: Deep learning architecture for TCGA sample differentiation. Illustrated is the multi-input deep neural network (DNN) with a concatenation layer. Each input channel is for the mRNA fold-change, copy number alterations, and protein fold-change. ReLu activation functions were used for all but the final layer, which was a sigmoid function. Binary cross-entropy was used for the loss function and ADAM as the optimizer.

4.2.4 Feature selection and importance

Shapley additive explanations (SHAP) values were determined using the SHAP python package [113] to define input contributions in histological classification in the deep models. This effectively assigns an importance to every input feature for a particular prediction on a sample. The scores aid in investigating which input channels (mRNA versus protein versus copy number) seem to drive predictions more regularly,

overall which genes seem to be differentiating and potentially important in the cancer histology, and if BDSGs fall within those groups.

After the SHAP values were calculated in the all features model, hierarchical clustering was performed using average linkage and Euclidean distances with the Python seaborn package on the fold changes and copy number data to determine if raw values could separate histological types. This was done for each of the input channels to reflect initial separation of the samples into their respective classes. It is important to note this methodology provides an estimation, as there are further dense layers after feature concatenation. For plotting distributions SHAP values were log-modulus transformed [114].

4.3 Results and Discussion

4.3.1 -omics signatures in differing cancer histologies

By integrating and joining a variety of genomic features across two cancer histologies one is presented with a unique opportunity. One can examine how patterns in expression and alteration influence protein expression and how these patterns can help differentiate cancer types. In BRCA there was a mRNA fold change of -0.0364 ± 0.9389 (mean and standard deviation), for OV 0.0581±1.004 and that they were significantly different (Mann-Whitney U, $P < 1 \times 10^{-16}$); CNV segment means for BRCA were 0.0216±0.318, for OV 0.0155±0.411 that that were significantly different (Mann-Whitney U, $P < 1 \times 10^{-16}$); and protein fold changes in BRCA -0.0003±0.9625, and OV 0.0005±0.6711 that were significantly different (Mann-Whitney U, $P < 3 \times 10^{-05}$). This high-level comparison supports the intuition that there are unique and differentiating transcriptional and translational behaviors in these histologies.

Figure 21 illustrates the correlational relationships between the breast cancer and ovarian cancer features. Each correlational distribution (e.g. mRNA vs Protein scores) were found to be significantly different between the two sample classes (two-sample KS, mRNA vs protein $P < 1.0 \times 10^{-16}$; mRNA vs CNV $P < 1.0x10^{-16}$, protein vs CNV $P < 1.0 \times 10^{-16}$). From visual inspection of the distributions, the differences in mRNA and copy number correlation is striking. In general, it appears that there is better correlation between copy number and mRNA changes within ovarian cancer samples. Breast cancer also has poorer correlation between mRNA fold changes and protein fold changes, suggesting some post transcriptional disruption or activity.

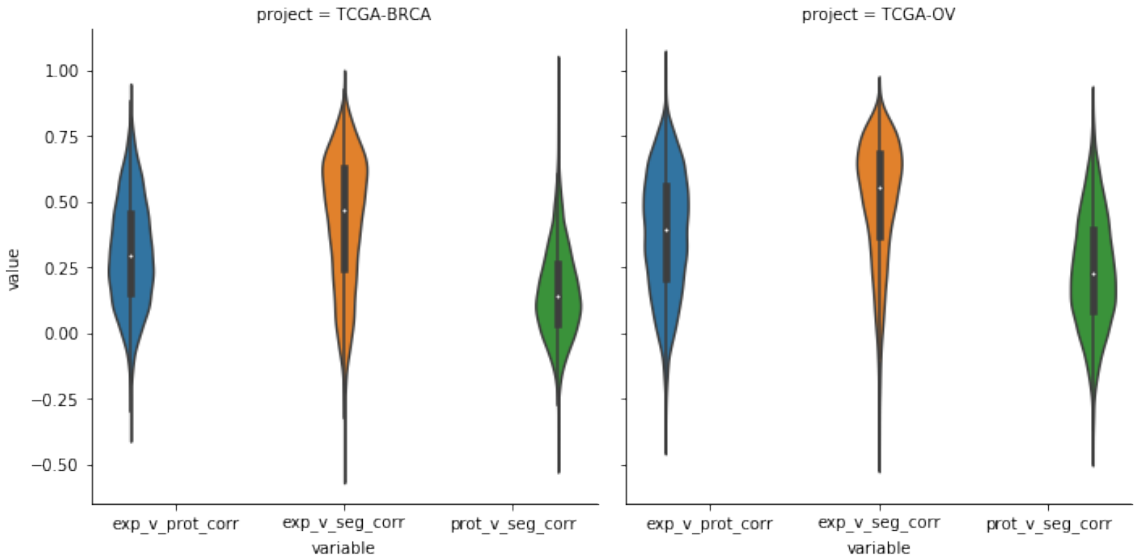


Figure 21: Correlation distributions for genomic features across BRCA and OV genes. This plot illustrates the correlation distributions between the three genomic features for both the BRCA and OV cohorts in this study. It emphasizes the transcriptional differences between BRCA and OV from the mRNA vs Protein and mRNA vs copy number differences.

I previously determined [109] 16 broadly dosage sensitive genes (BDSG) in BRCA. Using similar criteria 58 genes were identified from the OV cohort

(Supplemental Table 1). These counts are a surprising observation: the concordance of transcriptional activity seems more prevalent in OV samples.

Reviewing recent literature for the selected genes impact or relevance to cancer revealed their involvement in miRNA regulation or potential prognostic markers. miR-142-3p targets PPFIA1 and is implicated in HPV-induced tumorigenesis [115]. MIEN1 is a migration and invasion enhancer directly targeted by miR-136; suggesting it acts as a tumor suppressor and prognostic indicator in osteosarcoma [116,117]. APEX1 is targeted by miR-296-3p in non-small cell lung carcinomas and behaves like a tumor suppressor, inhibiting migration of cells [118]. The microRNA miR-193a-3p was found to slow the progress of HER2 positive breast cancer and was identified as a potential therapy candidate; however GRB7 over-expression seemed to counteract its ability to inhibit proliferation, migration, and infiltration of breast cancer cells [119]. Overexpression of Usp14 homologs in Chinese Hamster Ovary (CHO) cells had impacts on cell growth with miR-378-3p depletion [120]. miR-146a's expression impacted tumor growth and invasion through the VEGF/CDC42/PAK1 signaling pathway [121]. In two cases there was documentation that other regulatory elements were "sponging" miRNAs in the environment. Particularly BFAL1, a lncRNA, was downregulating miR-155-5p and miR-200a-3p and regulated RHEB expression. CBL.11, a circular RNA (circRNA) was regulating miR-6778-5p and effected YWHAE expression.

PPFIA1 overexpression levels in head and neck squamous cell carcinoma (HNSCC) and oropharyngeal squamous cell carcinoma (OPSCC) had associations with poor survival [122,123]. Increased metastatic relapse risk in estrogen receptor positive, nodal negative (ER+/N-) breast cancer could be indicated by PPFIA1 expression, which

was evidenced via KM analysis [124]. MIEN1 had evidence that positive protein production was indicative of a poorer survival rate [125]. Its copy number alterations and co-amplification with ERBB2 also highlighted it as a candidate cancer promoting gene with shorter overall survival [126]. USP32 was overexpressed in small cell lung carcinoma (SCLC) tissues and was highly correlated to disease stage and invasion; silencing it caused a decrease in cell proliferation and invasion [127]. Tumor proliferation and invasion in NSCLC was facilitated by TIMM50 through its modifications of the ERK/P90RSK signaling pathway [128].

All 74 BDSGs were selected out of both sample sets. BRCA BDSG mRNA fold change was 0.0474±0.964 (mean and standard deviation), OV -0.0757±1.049; BRCA copy number mean was 0.051±0.430, OV 0.063±0.500; and BRCA protein fold changes with a mean of -0.001±1.03, OV 0.001±0.929. In all three cases there were statistically significant differences between the two sample types when comparing the between the classes of genomic data (mRNA, Mann Whitney U, $P < 2.0 \times 10^{-12}$; CNV, Mann Whitney U, $P < 7 \times 10^{-9}$; protein, Mann Whitney U, $P = 0.034$). It is interesting to note that while statistically significant, the p-value for protein fold changes between the groups is orders of magnitude different from the remaining features.

Finally, the expression signatures of the separate BDSG classes were examined: 16 genes in the breast cancer samples and 58 in ovarian. For mRNA fold-change in expression the breast cancer BDSGs had a mean expression of 0.017±1.09 and ovarian samples were -0.088±1.102. The copy number means for BRCA were 0.186±0.647 and OV were 0.087±0.512; protein fold changes in BRCA were -0.109±1.09 and OV were -0.046±0.970. Of the histology-specific comparisons only mRNA and protein fold

changes were significantly different: Mann Whitney U, $P = 0.0139$ and $P = 1.13x10^{-7}$, respectively. Copy numbers were not significant different between the two groups of BDSGs. Figure 22 visualizes the results. Ultimately, while there are significant differences in the BDSG signatures there is not a clear trend suggesting uniform over- or underexpression. BDSGs do not appear to be a class of enhanced or suppressed genes.

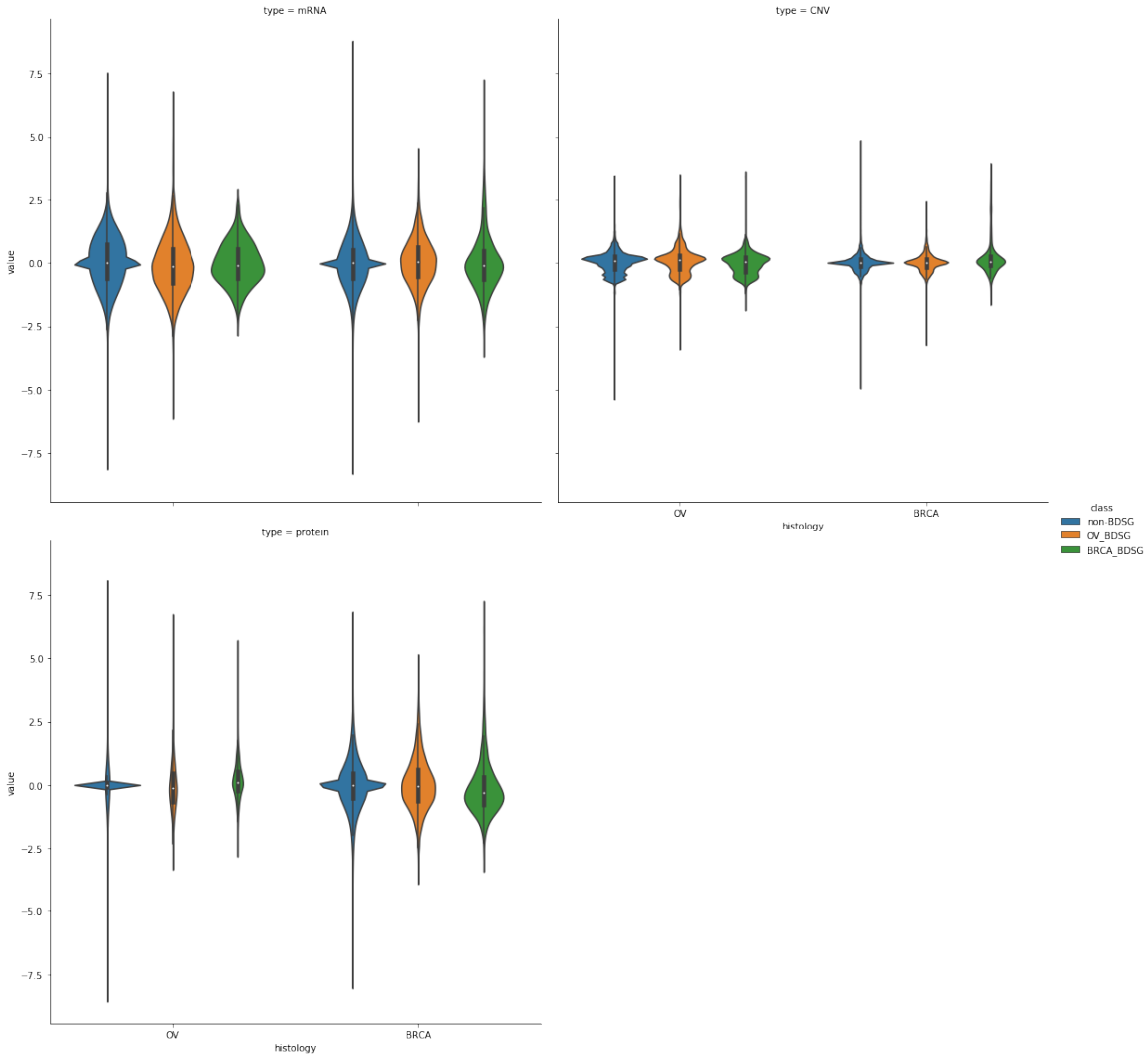


Figure 22: Expression distribution comparison between OV-, BRCA-, and non-BDSGs across genomic datatypes. The comparison of the values of the BDSG signatures in the genomics data illustrate differences in overall distributions between OV

and BRCA BDSGs. In cross-comparisons (e.g. OV BDSG from BRCA samples) there are differing distributions, but no distinct enhancement or repression.

It can be observed that in general the BRCA BDSGs show a more regular trend towards increased fold-changes, the exception being BRCA BDSGs in OV samples. In OV samples, the OV BDSGs reflect a more normal distribution in all but the protein fold changes. In BRCA samples, only in mRNA fold changes does one see a tendency towards down-regulation. This highlights the transcription dysregulation more common in breast cancers and coincides with observations in correlation signatures.

4.3.2 Machine learning approaches

The deep learning model had over 31,000 trainable parameters. Both the all feature deep learning model and BDSG-only model had 100% accuracy in predicting the tumor types of the validation set samples. Given the input size of the feature space versus the number of samples for the all feature model and the low loss and high accuracy, it is a safe assumption that the deep learning models are overfitted. In this case, that is satisfactory. The classification of histological cancer type acts largely as a validation metric for the model. While observing that protein, mRNA, and copy number alterations are unique to separate cancer types; there are much more pragmatic diagnosis methods. My interest is understanding which features drive this predictive process. To further determine prediction sensitivity to feature selection, with a focus on dosage-sensitivity (correlation of genomic features), non-BDSGs were selected at the same volume as BDSGs (n=74). These models could not differentiate samples. This helps suggests that the signal of BDSG genes within the all features model must still contribute in sample type identification.

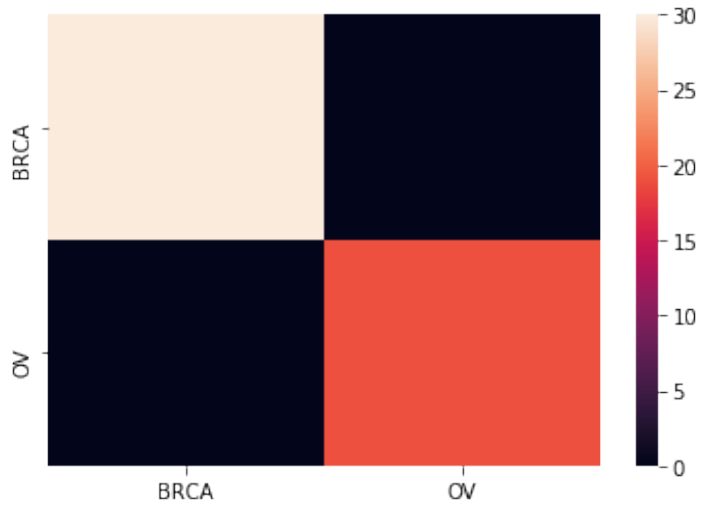


Figure 23: Deep learning test set prediction confusion matrix. Illustrated is the accurate prediction of sample types in the three-channel all gene deep neural network.

To confirm the utility of leveraging deep learning versus traditional methods; I compared predictions with a random forest model. I found that no matter the inputs provided to the random forest model, it would accurately label samples no matter the input size and input selection. The combination of information-content based methodologies and noise in the data seems to cause random forests to overfit, quickly. The random forest models also do not capture the nuanced interaction of the multiple genomic data types, nor can it offer a probabilistic score or regression value for the class assignment. Even when the input classes are shuffled, the Random Forest model has a moderately accurate prediction scheme. What remains is to determine the distribution of SHAP values within and between these two models and what classes of genes exist therein.

4.3.3 Feature importance in deep learning.

Briefly, SHAP values are a score based on information and game theory that provide an importance for a given feature on a prediction for a sample. By observing and

exploring the distributions of these values genes can be identified which aid in emphasizing the sample morphology and cancer significance. The sign of the value (positive or negative) indicates the "direction" of influence of the feature. For example, a negative value in our model indicates a shift towards being classified as a breast cancer sample; where positive is towards ovarian.

In the all features model there was an observed a mean absolute importance of $2.86 \times 10^{-5} \pm 3.50 \times 10^{-5}$ for mRNA features, $1.28 \times 10^{-6} \pm 1.94 \times 10^{-6}$ for copy number features, and $9.59 \times 10^{-6}\text{^} \pm 1.34 \times 10^{-5}$ for protein. When selecting for previously identified BDSGs out of this model I get the following: $2.55 \times 10^{-5} \pm 3.54\text{^}10^{-5}$ for mRNA, $1.72 \times 10^{-6} \pm 1.07 \times 10^{-5}$ for CNV, $1.07 \times 10^{-5} \pm 1.33 \times 10^{-5}$ for protein. The BDSG SHAP signatures across data types were significantly different from the remaining non-BDSGs (KS test, $P < 1 \times 10^{-16}$).

One can note how mRNA and protein features drive classification in the all features model based on the distribution of SHAP values (Figure 24, Supplemental Figure 1). When considering the mRNA SHAP distributions I can determine the differentiating nature of BDSGs. Within the OV samples BRCA BDSGs indicate a positive trend of training (towards an OV classification). This suggests that the expression signatures of these genes are unique and identifying and define the samples as 'not BRCA.' The same trend is observed within the BRCA cohort for OV BDSGs. If I had instead observed a positive trend for OV BDSGs in BRCA patients, this would indicate they were being incorrectly suggested as having ovarian cancer. The BDSGs were not necessarily the most important features to the model, however they clearly behaved in differentiating samples.

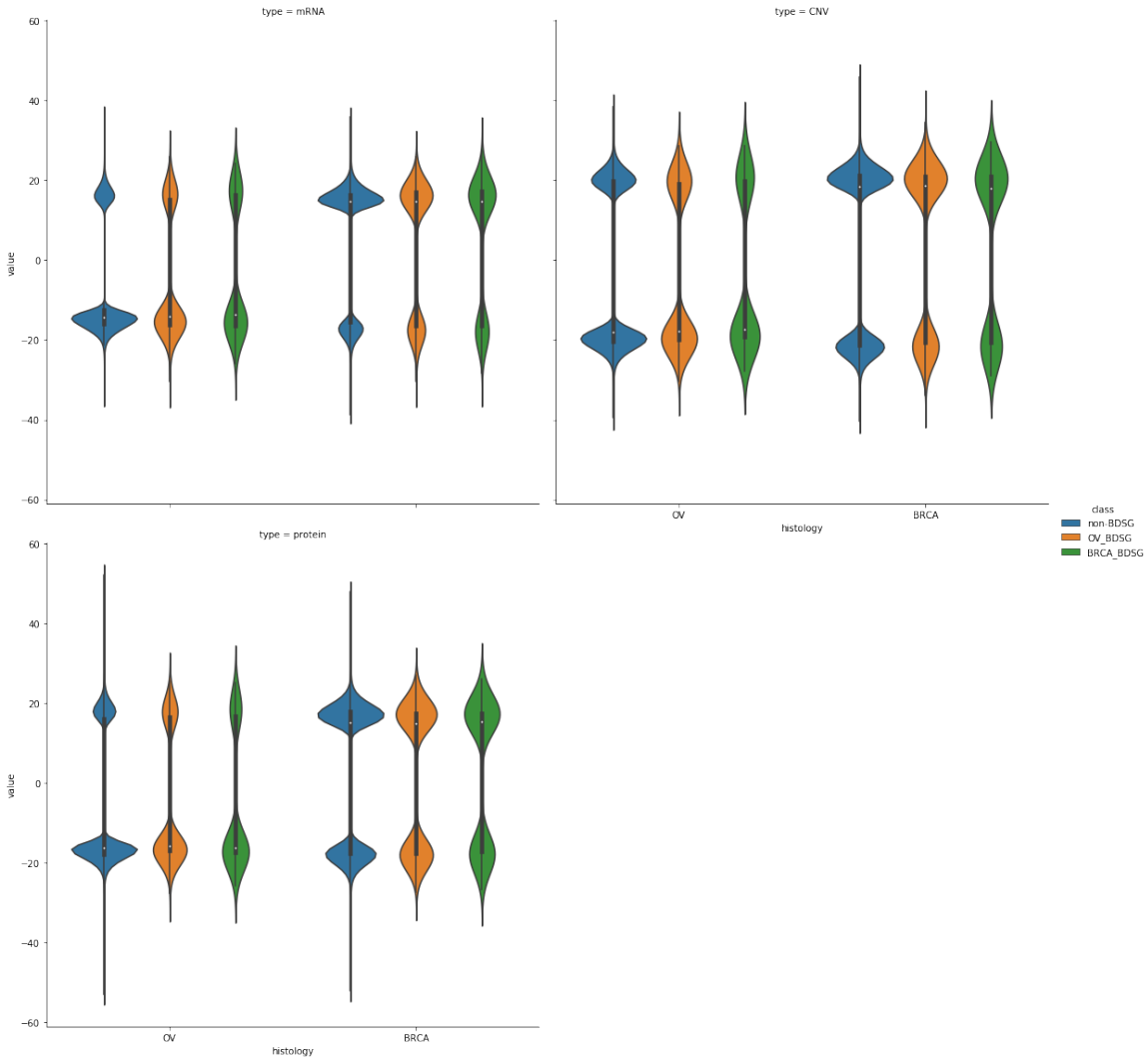


Figure 24: SHAP value distribution comparison between OV-, BRCA-, and non-BDSGs across genomic datatypes. Distributions of SHAP values emphasize their differential behavior. OV and BRCA BDSGs consistently identify sample types.

Previously I leveraged correlation across the three genetic data types to help do within-histology profiling. Performing such an exercise allows one to prioritize genes that may be effected by transcriptional disruption within a disease state. By leveraging the SHAP values from the deep learning model, I can cross-compare the cancers and suggest more subtle differences. By selecting the top 99th percentile of genes across the three input types, a set is generated that was used to perform gene set enrichment analysis

(GSEA) [129,130]. Doing so emphasizes two sets that are tied to early/late responses to estrogen ($P = 8.01 \times 10^{-7}; P = 1.14 \times 10^{-4}$; hypergeometric test), as well as genes regulated by the NF-kB in response to TNF ($P = 4.96 \times 10^{-8}$; hypergeometric test), and genes involved in p53 pathways and networks ($P = 2.99 \times 10^{-6}$; hypergeometric test). A number of immunological sets were also identified: genes upregulated by STAT5 in response to IL2 stimulation ($P = 2.99 \times 10^{-6}$; hypergeometric test), genes encoding components of the complement system, or innate immune system ($P = 2.04 \times 10^{-7}$; hypergeometric test), and genes upregulated during transplant rejection ($P = 2.99 \times 10^{-6}$; hypergeometric test). This suggests potential differences in immune cell responses or cell populations between the two cancer types; these differences may help in developing and understanding new immunotherapies [131]. Most importantly, these signatures could not have been detected with expression alone. The importance values were weakly correlated with their respective fold-change values (mRNA: 3.83×10^{-2}; protein: 5.93×10^{-2}; CNV: -2.57×10^{-1}) indicating that SHAP values are not purely a surrogate for differential expression. Performing the same analysis based off of fold-change expression emphasized much more general gene sets such as those involved with protein secretion ($P = 5.17 \times 10^{-8}$; hypergeometric test) or genes involved in homeostasis ($P = 6.45 \times 10^{-9}$; hypergeometric test).

Interestingly enough, deep learning models with randomly selected genes can still perform classification. The architecture of the model makes it sensitive to genes exhibiting tissue sensitivity, age differentiation, or other oncology related genes. This will be emphasized due to the low sample numbers for training and testing. In the case of

this study, I was limited to what is available in TCGA so inherent age and tissue-specific biases will exist within the dataset.

4.4 Conclusions

The transcriptional landscape of cancers continues to be a complex model. By examining these samples through a multi-genomics lens, we can begin to understand the mechanistic differences between these diseases. There continues to exist a unique subset of genes, which show matching changes in copy number, mRNA, and protein changes. These genes consistently differentiate the sample morphology, but overall are not the most important features for the classification. What remains to be understood are the functional implications of these subclasses.

Additionally, more data could be added to capture the regulatory changes that may occur. For example, miRNA or methylation data could help capture any alterations or interventions that may occur to inhibit expression. The investigations have also focused predominantly on genes with correlative behaviors. More complex signatures may identify other subclasses of genes that may be interesting. Genes, where copy number and mRNA changes correlate and protein poorly coincide, may illustrate interesting subjects of post-transcriptional modifications or regulation. Further explorations should be conducted to understand these patterns and their relationship to transcriptional regulators: long non-coding, microRNAs, RBPs, and transcription factors.

5. **Final Conclusions**

This body of work represents many years of synthesis, exploration, and analysis. Through its course, it has been structured to aid in expanding capabilities in data analysis and gathering, familiarizing with the oncogenic landscape, and incorporating bleeding edge technical implementations. It began with merely understanding transcriptional activity for important regulators and how they change in cancer. The project progressed to deepening the comprehension of transcriptional regulation in a specific cancer type and how those patterns identify genes, and finally, how these discrete patterns can emphasize critical functional differences between cancer types.

The initial work emphasized RBPs importance, given their consistently significant difference across a variety of tissues. They were found to be highly expressed in a variety of cancers, with a subset being many folds over-expressed in a majority of the cases. These strongly upregulated RBPs may potentially have interesting oncogenic roles, but cursory investigations based on protein interactions and network metrics did not readily identify a causative relationship. This study was initially a challenge because of the immaturity of the TCGA dataset at the time. Samples and cancer types had to be excluded because of missing data and at the time insufficient normal controls were available to use as a baseline. Given the time that has progressed since the initial investigation, it would be beneficial for others to attempt to expand on it with additional experimental and annotation data.

The second phase attempted to address how transcriptional changes in expression from healthy to disease states are reflected across the three cornerstone datatypes: protein, mRNA, and copy number alteration. The goal was to understand how

amplification impacted overall protein production and what the subsequent mRNA levels were like in the change from healthy to cancer states. The clear trend was the staggeringly poor correlation across the three features. Even more interesting was a small subset of genes that had a strong correlation. This small group included the well-documented oncogene ERBB2/HER2 and otherwise functionally unannotated genes. By comparing the effects of shRNA knockdown of these genes in cancer models, I illustrated how some of these relatively unknown genes behave similarly to known oncogenes and oncosupressors. The complexities of this study centered around the management and analysis of large volumes of data. While the final parsed and processed results were small, the initial files and their annotations were multiple gigabytes in size and required complex joins and pivots to prepare. If more mass spectra become available for other TCGA datasets, it would be beneficial to continue the exercise to build a library of correlational patterns and gene signatures.

The third and final phase attempts to compare cancer types based on the gene analysis in the second phase, and also leverage deep models to identify key genes that may explain the mechanism differences in the two cancers. From my previous work the impact of amplification and deletion is regulated to a large degree was identified in breast cancer. The observation was expanded to ovarian cancer, and immediately there was a larger proportion of highly correlative genes. It emphasized a fundamental difference in transcriptional or post-transcriptional regulation that was going on in the two cancer types. The deep learning model was trained to differentiate samples. Not necessarily as a classification exercise, but to leverage model explainability methods. These methods aided in the identification of the most important features, or genes, that differentiated the

samples. When compared to random forests or other machine learning approaches, this solution provided the most reliable classification and explainability. Even in comparison to the prioritization based on differential expression, deep learning prioritized features that identified gene sets pertinent to important, disrupted biological processes.

The major complication in working with deep learning models is data availability and feature/sample balancing. While there was a large volume of data on a given observation, there was not a large and varied sample set. This, of course, means my model and observations are very much tied to TCGA. In some cases, extra care had to be used to verify that the model was not arbitrarily classifying based on genes correlated to age or gender. Of course, like all the other efforts, when more samples are available with the necessary data, the model can only be further improved. Also, this work was done just to classify one cancer type or another. The dataset and approach can be expanded to build deep regressors to predict tumor growth, tumor size, or other features derived from actual tumor morphology.

6. **Appendix**

HGNC symbol	Chrom. No.	Karyotype band	Gene description
			aminoadipate-semialdehyde dehydrogenase-phosphopantetheinyl
AASDHPPT	11	q22.3	transferase
			ATP binding cassette
ABCF2	7	q36.1	subfamily F member 2
			ATP binding cassette
ABCF3	3	q27.1	subfamily F member 3
ACTN4	19	q13.2	actinin alpha 4
ADSS	1	q44	adenylosuccinate synthase
			AFG3 like matrix AAA
AFG3L2	18	p11.21	peptidase subunit 2
			apurinic/apyrimidinic
APEX1	14	q11.2	endodeoxyribonuclease 1
			ADP ribosylation factor guanine nucleotide
ARFGEF1	8	q13.2	exchange factor 1
			ADP ribosylation factor
ARFIP1	4	q31.3	interacting protein 1

BROX	1	q41	BRO1 domain and CAAX motif containing
CAP1	1	p34.2	cyclase associated actin cytoskeleton regulatory protein 1
CDC37	19	p13.2	cell division cycle 37
CHORDC1	11	q14.3	cysteine and histidine rich domain containing 1
CTPS1	1	p34.2	CTP synthase 1
DLAT	11	q23.1	dihydrolipoamide S-acetyltransferase
DNM2	19	p13.2	dynamin 2
ECH1	19	q13.2	enoyl-CoA hydratase 1
EIF3J	15	q21.1	eukaryotic translation initiation factor 3 subunit J
FLOT2	17	q11.2	flotillin 2
GARS	7	p14.3	glycyl-tRNA synthetase
GET3	19	p13.13	guided entry of tail-anchored proteins factor 3, ATPase

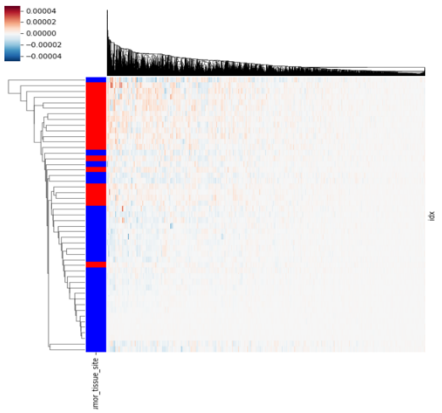
			glyoxalase domain
GLOD4	17	p13.3	containing 4
HARS	5	q31.3	histidyl-tRNA synthetase
			heterogeneous nuclear
HNRNPUL1	19	q13.2	ribonucleoprotein U like 1
IPO5	13	q32.2	importin 5
KARS	16	q23.1	lysyl-tRNA synthetase
			LPS responsive beige-like
LRBA	4	q31.3	anchor protein
			microtubule actin
MACF1	1	p34.3	crosslinking factor 1
			methylthioadenosine
MTAP	9	p21.3	phosphorylase
MTDH	8	q22.1	metadherin
NSFL1C	20	p13	NSFL1 cofactor
			OPA1 mitochondrial
OPA1	3	q29	dynamin like GTPase
			platelet activating factor
			acetylhydrolase 1b
PAFAH1B1	17	p13.3	regulatory subunit 1

			p21 (RAC1) activated kinase
PAK1	11	q14.1	1
			p21 (RAC1) activated kinase
PAK4	19	q13.2	4
			programmed cell death 6
PDCD6IP	3	p22.3	interacting protein
PGLS	19	p13.11	6-phosphogluconolactonase
			phospholipase A2 activating
PLAA	9	p21.2	protein
			pyridoxal phosphate binding
PLPBP	8	p11.23	protein
PTK2	8	q24.3	protein tyrosine kinase 2
			pyrroline-5-carboxylate
PYCR3	8	q24.3	reductase 3
			Ras homolog, mTORC1
RHEB	7	q36.1	binding
			seryl-tRNA synthetase 2,
SARS2	19	q13.2	mitochondrial
			SGT1 homolog, MIS12
			kinetochore complex
SUGT1	13	q14.3	assembly cochaperone

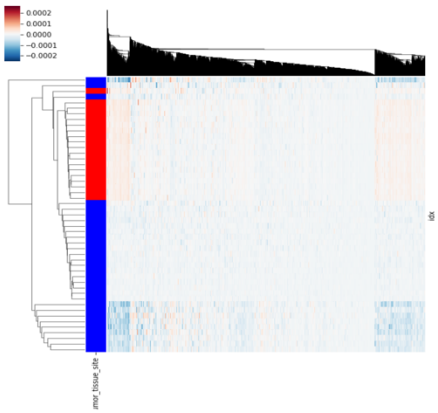
			SPT5 homolog, DSIF
SUPT5H	19	q13.2	elongation factor subunit
TBCB	19	q13.12	tubulin folding cofactor B
			translocase of inner
			mitochondrial membrane
TIMM44	19	p13.2	44
			translocase of inner
			mitochondrial membrane
TIMM50	19	q13.2	50
			tripartite motif containing
TRIM33	1	p13.2	33
			Ts translation elongation
TSFM	12	q14.1	factor, mitochondrial
			ubiquitin conjugating
UBE2L3	22	q11.21	enzyme E2 L3
UFL1	6	q16.1	UFM1 specific ligase 1
			USO1 vesicle transport
USO1	4	q21.1	factor
			ubiquitin specific peptidase
USP14	18	p11.32	14
VCP	9	p13.3	valosin containing protein

VPS26B	11	q25	VPS26, retromer complex component B
YWHAE	17	p13.3	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein epsilon
ZMPSTE24	1	p34.2	zinc metallopeptidase STE24

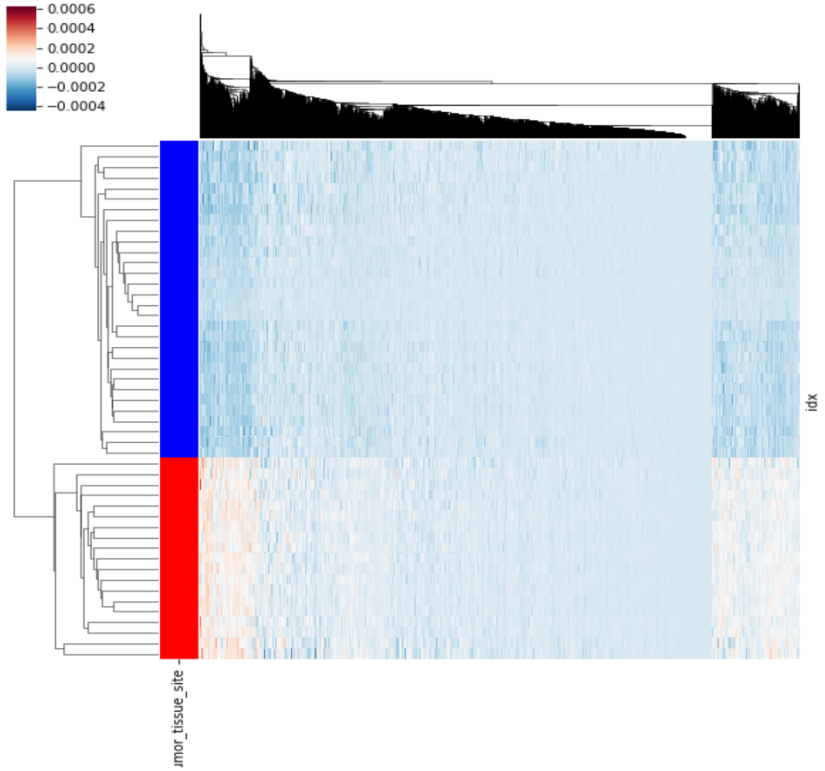
Supplemental Table 1: **Ovarian cancer identified BDSGs, their chromosomal position, and description.**



(a)



(b)



(c)

Supplemental Figure 1: Heat map and hierarchical clustering of SHAP values for all gene deep neural network. By clustering SHAP (columns: genes, rows: samples) values we can estimate 'decisions' by the model across the three input feature sets to pedict

cancer type (red: OV, blue: BRCA): (a) CNV channel, (b) Protein channel, and (c) mRNA channel. The clustering emphasizes the protein and mRNA contributions through the separation of the two groups.

5. **References**

1. Glisovic T, Bachorik JL, Yong J, Dreyfuss G. RNA-binding proteins and post-transcriptional gene regulation. FEBS Lett. 2008/03/13. 2008;582: 1977–1986. doi:10.1016/j.febslet.2008.03.004
2. Keene JD. RNA regulons: coordination of post-transcriptional events. Nat Rev Genet. 2007;8: 533–543. doi:10.1038/nrg2111
3. Janga SC. From specific to global analysis of posttranscriptional regulation in eukaryotes: posttranscriptional regulatory networks. Brief Funct Genomics. 2012;11: 505–521. doi:10.1093/bfgp/els046
4. Lukong KE, Chang K, Khandjian EW, Richard S. RNA-binding proteins in human genetic disease. Trends Genet. Elsevier; 2008;24: 416–425. doi:10.1016/j.tig.2008.05.004
5. Kim MY, Jeong* JH & S. <p>Emerging roles of RNA and RNA-binding protein network in cancer cells</p> BMB Rep. 2009/03/31. Korean Society for Biochemistry and Molecular Biology; 2009;42: 125–130. Available: http://www.bmbreprots.org/journal/view.html?doi=
6. Castello A, Fischer B, Hentze MW, Preiss T. RNA-binding proteins in Mendelian disease. Trends Genet. Elsevier; 2013;29: 318–327. doi:10.1016/j.tig.2013.01.004
7. Musunuru K. Cell-Specific RNA-Binding Proteins in Human Disease. Trends Cardiovasc Med. 2003;13: 188–195. doi:https://doi.org/10.1016/S1050-1738(03)00075-6
8. Mittal N, Roy N, Babu MM, Janga SC. Dissecting the expression dynamics of RNA-binding proteins in posttranscriptional regulatory networks. Proc Natl Acad

Sci U S A. 2009/11/16. National Academy of Sciences; 2009;106: 20300–20305. doi:10.1073/pnas.0906940106

9. Mittal N, Scherrer T, Gerber AP, Janga SC. Interplay between Posttranscriptional and Posttranslational Interactions of RNA-Binding Proteins. J Mol Biol. 2011;409: 466–479. doi:https://doi.org/10.1016/j.jmb.2011.03.064

10. Wurth L. Versatility of RNA-Binding Proteins in Cancer. Comp Funct Genomics. 2012/05/14. Hindawi Publishing Corporation; 2012;2012: 178525. doi:10.1155/2012/178525

11. Lima L, Morais A, Lobo F, Calais-da-Silva FM, Calais-da-Silva FE, Medeiros R. Association between FAS polymorphism and prostate cancer development. Prostate Cancer Prostatic Dis. 2008;11: 94–98. doi:10.1038/sj.pcan.4501002

12. Izquierdo JM. Hu Antigen R (HuR) Functions as an Alternative Pre-mRNA Splicing Regulator of Fas Apoptosis-promoting Receptor on Exon Definition. J Biol Chem . 2008;283: 19077–19084. doi:10.1074/jbc.M800017200

13. Izquierdo JM. Cell-specific regulation of Fas exon 6 splicing mediated by Hu antigen R. Biochem Biophys Res Commun. 2010;402: 324–328. doi:https://doi.org/10.1016/j.bbrc.2010.10.025

14. Izquierdo JM, Majós N, Bonnal S, Martínez C, Castelo R, Guigó R, et al. Regulation of Fas Alternative Splicing by Antagonistic Effects of TIA-1 and PTB on Exon Definition. Mol Cell. Elsevier; 2005;19: 475–484. doi:10.1016/j.molcel.2005.06.015

15. Lee M, Dworkin AM, Gildea D, Trivedi NS, Program NCS, Moorhead GB, et al. RRP1B is a metastasis modifier that regulates the expression of alternative mRNA

isoforms through interactions with SRSF1. Oncogene. 2013/04/22. 2014;33: 1818–1827. doi:10.1038/onc.2013.133

16. Mayr C, Bartel DP. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell. 2009;138: 673–684. doi:10.1016/j.cell.2009.06.016

17. Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, et al. Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. Cell. 149: 1393–1406. doi:10.1016/j.cell.2012.04.031

18. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Res. 2012;22: 1775–1789. doi:10.1101/gr.132159.111

19. National Cancer Institute, National Human Genome Research Institute. The Cancer Genome Atlas [Internet]. 2010 [cited 1 Jan 2015]. Available: http://cancergenome.nih.gov/

20. Rustici G, Kolesnikov N, Brandizi M, Burdett T, Dylag M, Emam I, et al. ArrayExpress update--trends in database growth and links to data analysis tools. Nucleic Acids Res. 2012/11/27. Oxford University Press; 2013;41: D987–D990. doi:10.1093/nar/gks1174

21. Human BodyMap (HBM) [Internet].

22. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. Cold Spring Harbor Laboratory Press; 2012;22: 1760–1774.

doi:10.1101/gr.135350.111

23. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. 2010/10/30. Oxford University Press; 2011;39: D152–D157. doi:10.1093/nar/gkq1027

24. Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA. DBD--taxonomically broad transcription factor predictions: new content and functionality. Nucleic Acids Res. 2007/12/11. Oxford University Press; 2008;36: D88–D92. doi:10.1093/nar/gkm964

25. Chatr-aryamontri A, Breitkreutz B-J, Oughtred R, Boucher L, Heinicke S, Chen D, et al. The BioGRID interaction database: 2015 update. Nucleic Acids Res. 2015;43: D470--D478. doi:10.1093/nar/gku1204

26. Csardi G, Nepusz T. The Igraph Software Package for Complex Network Research. InterJournal. 2005;Complex Sy: 1695.

27. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, et al. CORUM: the comprehensive resource of mammalian protein complexes--2009. Nucleic Acids Res. 2009/11/01. Oxford University Press; 2010;38: D497–D501. doi:10.1093/nar/gkp914

28. Győrffy B, Lanczky A, Eklund AC, Denkert C, Budczies J, Li Q, et al. An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. Breast Cancer Res Treat. 2010;123: 725–731. doi:10.1007/s10549-009-0674-9

29. Darnell RB. HITS-CLIP: panoramic views of protein-RNA regulation in living cells. Wiley Interdiscip Rev RNA. 2010/08/02. 2010;1: 266–286.

doi:10.1002/wrna.31

30. Cooper TA, Wan L, Dreyfuss G. RNA and disease. Cell. 2009;136: 777–793. doi:10.1016/j.cell.2009.02.011

31. Lui W-Y, Cheng CY. Transcriptional regulation of cell adhesion at the blood-testis barrier and spermatogenesis in the testis. Adv Exp Med Biol. 2012;763: 281–294. doi:10.1007/978-1-4614-4711-5_14

32. Green SM, Mostaghel EA, Nelson PS. Androgen action and metabolism in prostate cancer. Mol Cell Endocrinol. 2012/03/20. 2012;360: 3–13. doi:10.1016/j.mce.2011.09.046

33. Baena E, Shao Z, Linn DE, Glass K, Hamblen MJ, Fujiwara Y, et al. ETV1 directs androgen metabolism and confers aggressive prostate cancer in targeted mice and patients. Genes Dev. Cold Spring Harbor Laboratory Press; 2013;27: 683–698. doi:10.1101/gad.211011.112

34. Aragaki M, Takahashi K, Akiyama H, Tsuchiya E, Kondo S, Nakamura Y, et al. Characterization of a Cleavage Stimulation Factor, 3′ pre-RNA, Subunit 2, 64 kDa (CSTF2) as a Therapeutic Target for Lung Cancer. Clin Cancer Res. 2011;17: 5889 LP – 5900. doi:10.1158/1078-0432.CCR-11-0240

35. Cruciat C-M, Dolde C, de Groot REA, Ohkawara B, Reinhard C, Korswagen HC, et al. RNA Helicase DDX3 Is a Regulatory Subunit of Casein Kinase 1 in Wnt–β-Catenin Signaling. Science (80-). 2013;339: 1436 LP – 1441. doi:10.1126/science.1231499

36. Botlagunta M, Vesuna F, Mironchik Y, Raman A, Lisok A, Winnard Jr P, et al. Oncogenic role of DDX3 in breast cancer biogenesis. Oncogene. 2008/02/11.

2008;27: 3912–3922. doi:10.1038/onc.2008.33

37. Wu D-W, Liu W-S, Wang J, Chen C-Y, Cheng Y-W, Lee H. Reduced p21^{WAF1/CIP1} via Alteration of p53-DDX3 Pathway Is Associated with Poor Relapse-Free Survival in Early-Stage Human Papillomavirus–Associated Lung Cancer. Clin Cancer Res. 2011;17: 1895 LP – 1905. doi:10.1158/1078-0432.CCR-10-2316

38. Alawi F, Lin P. Dyskerin is required for tumor cell growth through mechanisms that are independent of its role in telomerase and only partially related to its function in precursor rRNA processing. Mol Carcinog. 2010/12/10. 2011;50: 334–345. doi:10.1002/mc.20715

39. Katunaric M, Zamolo G. Modulating telomerase activity in tumor patients by targeting dyskerin binding site for hTR. Med Hypotheses. 2012;79: 319–320. doi:https://doi.org/10.1016/j.mehy.2012.05.021

40. Liu B, Zhang J, Huang C, Liu H. Dyskerin overexpression in human hepatocellular carcinoma is associated with advanced clinical stage and poor patient prognosis. PLoS One. 2012/08/13. Public Library of Science; 2012;7: e43147–e43147. doi:10.1371/journal.pone.0043147

41. Bedolla RG, Wang Y, Asuncion A, Chamie K, Siddiqui S, Mudryj MM, et al. Nuclear versus cytoplasmic localization of filamin A in prostate cancer: immunohistochemical correlation with metastases. Clin Cancer Res. 2009;15: 788–796. doi:10.1158/1078-0432.CCR-08-1402

42. URAMOTO H, AKYÜREK LM, HANAGIRI T. A Positive Relationship Between Filamin and VEGF in Patients with Lung Cancer. Anticancer Res . 2010;30: 3939–

3944. Available: http://ar.iiarjournals.org/content/30/10/3939.abstract

43. Ai J, Huang H, Lv X, Tang Z, Chen M, Chen T, et al. FLNA and PGK1 are Two Potential Markers for Progression in Hepatocellular Carcinoma. Cell Physiol Biochem. 2011;27: 207–216. doi:10.1159/000327946

44. Nallapalli RK, Ibrahim MX, Zhou AX, Bandaru S, Sunkara SN, Redfors B, et al. Targeting filamin A reduces K-RAS-induced lung adenocarcinomas and endothelial response to tumor growth in mice. Mol Cancer. BioMed Central; 2012;11: 50. doi:10.1186/1476-4598-11-50

45. Okamoto N, Yasukawa M, Nguyen C, Kasim V, Maida Y, Possemato R, et al. Maintenance of tumor initiating cells of defined genetic composition by nucleostemin. Proc Natl Acad Sci U S A. 2011/07/05. National Academy of Sciences; 2011;108: 20388–20393. doi:10.1073/pnas.1015171108

46. Rao MRKS, Kumari G, Balasundaram D, Sankaranarayanan R, Mahalingam S. A Novel Lysine-rich Domain and GTP Binding Motifs Regulate the Nucleolar Retention of Human Guanine Nucleotide Binding Protein, GNL3L. J Mol Biol. 2006;364: 637–654. doi:https://doi.org/10.1016/j.jmb.2006.09.007

47. Kurokawa M, Kim J, Geradts J, Matsuura K, Liu L, Ran X, et al. A network of substrates of the E3 ubiquitin ligases MDM2 and HUWE1 control apoptosis independently of p53. Sci Signal. 2013;6: ra32–ra32. doi:10.1126/scisignal.2003741

48. Lu Z, Li Y, Takwi A, Li B, Zhang J, Conklin DJ, et al. miR-301a as an NF-κB activator in pancreatic cancer cells. EMBO J. 2010/11/26. Nature Publishing Group; 2011;30: 57–67. doi:10.1038/emboj.2010.296

49. Krietsch J, Caron M-C, Gagné J-P, Ethier C, Vignard J, Vincent M, et al. PARP activation regulates the RNA-binding protein NONO in the DNA damage response to DNA double-strand breaks. Nucleic Acids Res. 2012/08/31. Oxford University Press; 2012;40: 10287–10301. doi:10.1093/nar/gks798

50. Tsofack SP, Garand C, Sereduk C, Chow D, Aziz M, Guay D, et al. NONO and RALY proteins are required for YB-1 oxaliplatin induced resistance in colon adenocarcinoma cell lines. Mol Cancer. BioMed Central; 2011;10: 145. doi:10.1186/1476-4598-10-145

51. Van Vlierberghe P, Palomero T, Khiabanian H, Van der Meulen J, Castillo M, Van Roy N, et al. PHF6 mutations in T-cell acute lymphoblastic leukemia. Nat Genet. 2010/03/14. 2010;42: 338–342. doi:10.1038/ng.542

52. Yoo NJ, Kim YR, Lee SH. Somatic mutation of PHF6 gene in T-cell acute lymphoblatic leukemia, acute myelogenous leukemia and hepatocellular carcinoma. Acta Oncol (Madr). Taylor & Francis; 2012;51: 107–111. doi:10.3109/0284186X.2011.592148

53. Wang J, Leung JW, Gong Z, Feng L, Shi X, Chen J. PHF6 regulates cell cycle progression by suppressing ribosomal RNA synthesis. J Biol Chem. 2012/12/10. American Society for Biochemistry and Molecular Biology; 2013;288: 3174–3183. doi:10.1074/jbc.M112.414839

54. Ehlén Å, Brennan DJ, Nodin B, O'Connor DP, Eberhard J, Alvarado-Kristensson M, et al. Expression of the RNA-binding protein RBM3 is associated with a favourable prognosis and cisplatin sensitivity in epithelial ovarian cancer. J Transl Med. 2010;8: 78. doi:10.1186/1479-5876-8-78

55. Jögi A, Brennan DJ, Rydén L, Magnusson K, Fernö M, Stål O, et al. Nuclear expression of the RNA-binding protein RBM3 is associated with an improved clinical outcome in breast cancer. Mod Pathol. United States and Canadian Academy of Pathology, Inc.; 2009;22: 1564. Available: https://doi.org/10.1038/modpathol.2009.124

56. Jonsson L, Bergman J, Nodin B, Manjer J, Pontén F, Uhlén M, et al. Low RBM3 protein expression correlates with tumour progression and poor prognosis in malignant melanoma: an analysis of 215 cases from the Malmö Diet and Cancer Study. J Transl Med. BioMed Central; 2011;9: 114. doi:10.1186/1479-5876-9-114

57. Zeng Y, Wodzenski D, Gao D, Shiraishi T, Terada N, Li Y, et al. Stress-Response Protein RBM3 Attenuates the Stem-like Properties of Prostate Cancer Cells by Interfering with CD44 Variant Splicing. Cancer Res. 2013;73: 4123 LP – 4133. doi:10.1158/0008-5472.CAN-12-1343

58. Adamson B, Smogorzewska A, Sigoillot FD, King RW, Elledge SJ. A genome-wide homologous recombination screen identifies the RNA-binding protein RBMX as a component of the DNA-damage response. Nat Cell Biol. 2012;14: 318–328. doi:10.1038/ncb2426

59. Moudry P, Lukas C, Macurek L, Hanzlikova H, Hodny Z, Lukas J, et al. Ubiquitin-activating enzyme UBA1 is required for cellular response to DNA damage. Cell Cycle. Taylor & Francis; 2012;11: 1573–1582. doi:10.4161/cc.19978

60. Xu GW, Ali M, Wood TE, Wong D, Maclean N, Wang X, et al. The ubiquitin-activating enzyme E1 as a therapeutic target for the treatment of leukemia and multiple myeloma. Blood. 2010/01/14. American Society of Hematology;

2010;115: 2251–2259. doi:10.1182/blood-2009-07-231191

61. Kashiwaya K, Nakagawa H, Hosokawa M, Mochizuki Y, Ueda K, Piao L, et al. Involvement of the Tubulin Tyrosine Ligase-Like Family Member 4 Polyglutamylase in PELP1 Polyglutamylation and Chromatin Remodeling in Pancreatic Cancer Cells. Cancer Res. 2010;70: 4024 LP – 4033. doi:10.1158/0008-5472.CAN-09-4444

62. Bernassola F, Karin M, Ciechanover A, Melino G. The HECT Family of E3 Ubiquitin Ligases: Multiple Players in Cancer Development. Cancer Cell. Elsevier; 2008;14: 10–21. doi:10.1016/j.ccr.2008.06.001

63. Kreft SG, Nassal M. hRUL138, a novel human RNA-binding RING-H2 ubiquitin-protein ligase. J Cell Sci. 2003;116: 605 LP – 616. doi:10.1242/jcs.00261

64. Cano F, Miranda-Saavedra D, Lehner PJ. RNA-binding E3 ubiquitin ligases: novel players in nucleic acid regulation. Biochem Soc Trans. 2010;38: 1621–1626. doi:10.1042/BST0381621

65. Scherrer T, Mittal N, Janga SC, Gerber AP. A Screen for RNA-Binding Proteins in Yeast Indicates Dual Functions for Many Enzymes. PLoS One. Public Library of Science; 2010;5: e15499. Available: https://doi.org/10.1371/journal.pone.0015499

66. Lower KM, Turner G, Kerr BA, Mathews KD, Shaw MA, Gedeon ÁK, et al. Mutations in PHF6 are associated with Börjeson–Forssman –Lehmann syndrome. Nat Genet. 2002;32: 661–665. doi:10.1038/ng1040

67. BÖRJESON M, FORSSMAN H, LEHMANN O. An X-linked, Recessively Inherited Syndrome Characterized by Grave Mental Deficiency, Epilepsy, and Endocrine Disorder. Acta Med Scand. John Wiley & Sons, Ltd (10.1111);

1962;171: 13–22. doi:10.1111/j.0954-6820.1962.tb04162.x

68. Turner G, Lower KM, White SM, Delatycki M, Lampe AK, Wright M, et al. The clinical picture of the Börjeson–Forssman–Lehmann syndrome in males and heterozygous females with PHF6 mutations. Clin Genet. 2004;65: 226–232. doi:doi:10.1111/j.0009-9163.2004.00215.x

69. Deng W, Lopez-Camacho C, Tang J-Y, Mendoza-Villanueva D, Maya-Mendoza A, Jackson DA, et al. Cytoskeletal protein filamin A is a nucleolar protein that suppresses ribosomal RNA gene transcription. Proc Natl Acad Sci U S A. 2012/01/17. National Academy of Sciences; 2012;109: 1524–1529. doi:10.1073/pnas.1107879109

70. Cieśla J. Metabolic enzymes that bind RNA: Yet another level of cellular regulatory network? Acta Biochim Pol. 2006;53: 11–32.

71. Tsvetanova NG, Klass DM, Salzman J, Brown PO. Proteome-Wide Search Reveals Unexpected RNA-Binding Proteins in Saccharomyces cerevisiae. PLoS One. Public Library of Science; 2010;5: e12671. Available: https://doi.org/10.1371/journal.pone.0012671

72. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4: 44–57. doi:10.1038/nprot.2008.211

73. König J, Zarnack K, Luscombe NM, Ule J. Protein–RNA interactions: new genomic technologies and perspectives. Nat Rev Genet. 2012;13: 77–83. doi:10.1038/nrg3141

74. Ray D, Kazan H, Chan ET, Castillo LP, Chaudhry S, Talukder S, et al. Rapid and

systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nat Biotechnol. 2009;27: 667–670. doi:10.1038/nbt.1550

75. Gordon DJ, Resio B, Pellman D. Causes and consequences of aneuploidy in cancer. Nat Rev Genet. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2012;13: 189. Available: http://dx.doi.org/10.1038/nrg3123

76. Giam M, Rancati G. Aneuploidy and chromosomal instability in cancer: a jackpot to chaos. Cell Div. 2015;10: 3. doi:10.1186/s13008-015-0009-7

77. Santaguida S, Amon A. Short- and long-term effects of chromosome mis-segregation and aneuploidy. Nat Rev Mol Cell Biol. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2015;16: 473. Available: http://dx.doi.org/10.1038/nrm4025

78. Bloomfield M, Duesberg P. Inherent variability of cancer-specific aneuploidy generates metastases. Molecular Cytogenetics. London; 2016. doi:10.1186/s13039-016-0297-x

79. O'Connor C. Chromosomal Abnormalities: Aneuploidies. Nat Educ. 2008;1: 172.

80. Boveri T. Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. J Cell Sci. The Company of Biologists Ltd; 2008; 1–84. doi:10.1242/jcs.025742

81. Parris TZ, Kovacs A, Hajizadeh S, Nemes S, Semaan M, Levin M, et al. Frequent MYC coamplification and DNA hypomethylation of multiple genes on 8q in 8p11-p12-amplified breast carcinomas. Oncogenesis. 2014;3: e95. doi:10.1038/oncsis.2014.8

82. Tabach Y, Kogan-Sakin I, Buganim Y, Solomon H, Goldfinger N, Hovland R, et al. Amplification of the 20q chromosomal arm occurs early in tumorigenic transformation and may initiate cancer. PLoS One. 2011;6: e14632. doi:10.1371/journal.pone.0014632

83. Torres EM, Springer M, Amon A. No current evidence for widespread dosage compensation in S. cerevisiae. Odom DT, editor. Elife. eLife Sciences Publications, Ltd; 2016;5: e10996. doi:10.7554/eLife.10996

84. Dumaual CM, Steere BA, Walls CD, Wang M, Zhang Z-Y, Randall SK. Integrated Analysis of Global mRNA and Protein Expression Data in HEK293 Cells Overexpressing PRL-1. PLoS One. 2013;8: e72977. doi:10.1371/journal.pone.0072977

85. Greenbaum D, Colangelo C, Williams K, Gerstein M. Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biol. 2003;4: 117. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC193646/

86. Maier T, Güell M, Serrano L. Correlation of mRNA and protein in complex biological samples. FEBS Lett. 2009;583: 3966–3973. doi:http://dx.doi.org/10.1016/j.febslet.2009.10.036

87. Tuller T, Kupiec M, Ruppin E. Determinants of Protein Abundance and Translation Efficiency in S. cerevisiae. PLoS Comput Biol. 2007;3: e248. doi:10.1371/journal.pcbi.0030248

88. Vogel C, de Sousa Abreu R, Ko D, Le S, Shapiro BA, Burns SC, et al. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. Mol Syst Biol. 2010;6. doi:10.1038/msb.2010.59

89. Cancer Facts & Figures 2015 [Internet]. 2015. Available: http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2015/index

90. Tibes R, Qiu Y, Lu Y, Hennessy B, Andreeff M, Mills GB, et al. Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. Mol Cancer Ther. 2006;5: 2512–2521. doi:10.1158/1535-7163.mct-06-0334

91. Ellis MJ, Gillette M, Carr SA, Paulovich AG, Smith RD, Rodland KK, et al. Connecting Genomic Alterations to Cancer Biology with Proteomics: The NCI Clinical Proteomic Tumor Analysis Consortium. Cancer Discov. 2013;3: 1108–1112. doi:10.1158/2159-8290.CD-13-0219

92. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2015;43: D805--D811. doi:10.1093/nar/gku1075

93. Cowley GS, Weir BA, Vazquez F, Tamayo P, Scott JA, Rusin S, et al. Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. Sci Data. The Author(s); 2014;1: 140035. Available: http://dx.doi.org/10.1038/sdata.2014.35

94. Edwards NJ, Markey SP, Stein SE. Protein Reports – CPTAC Common Data Analysis Pipeline (CDAP) [Internet]. 2016. Available: https://cptac-data-portal.georgetown.edu/cptac/documents/CDAP_ProteinReports_description_20160503.pdf

95. Reiter L, Claassen M, Schrimpf SP, Jovanovic M, Schmidt A, Buhmann JM, et al. Protein Identification False Discovery Rates for Very Large Proteomics Data Sets

Generated by Tandem Mass Spectrometry. Mol & Cell Proteomics. 2009;8: 2405 LP – 2417. Available: http://www.mcponline.org/content/8/11/2405.abstract

96. Laddha S V, Ganesan S, Chan CS, White E. Mutational Landscape of the Essential Autophagy Gene BECN1in Human Cancers . Mol Cancer Res. 2014;12: 485–490. doi:10.1158/1541-7786.mcr-13-0614

97. Technologies N. Reference Genes for Normalization of Expression Data [Internet]. 2009 [cited 1 Jan 2016]. Available: https://www.nanostring.com/application/files/7014/8943/0117/TN_Normalization_of_Expression_Data.pdf

98. Liu Y, Beyer A, Aebersold R. On the Dependency of Cellular Protein Levels on mRNA Abundance. Cell. 2016. pp. 535–550. doi:10.1016/j.cell.2016.03.014

99. Garcia-Murillas I, Sharpe R, Pearson A, Campbell J, Natrajan R, Ashworth A, et al. An siRNA screen identifies the GNAS locus as a driver in 20q amplified breast cancer. Oncogene. Nature Publishing Group; 2014;33: 2478–2486. doi:10.1038/onc.2013.202

100. Rasool N, LaRochelle W, Zhong H, Ara G, Cohen J, Kohn EC. Secretory Leukocyte Protease Inhibitor, SLPI, Antagonizes Paclitaxel in Ovarian Cancer Cells. Clin Cancer Res. 2010;16: 600–609. doi:10.1158/1078-0432.CCR-09-1979

101. Tanaka S, Mori M, Akiyoshi T, Tanaka Y, Mafune K, Wands JR, et al. Coexpression of Grb7 with Epidermal Growth Factor Receptor or Her2/erbB2 in Human Advanced Esophageal Carcinoma. Cancer Res. 1997;57: 28. Available:

http://cancerres.aacrjournals.org/content/57/1/28.abstract

102. Slattery ML, Lundgreen A, Herrick JS, Wolff RK. Genetic variation in RPS6KA1, RPS6KA2, RPS6KB1, RPS6KB2, and PDK1 and risk of colon or rectal cancer. Mutat Res Mol Mech Mutagen. 2011;706: 13–20. doi:http://dx.doi.org/10.1016/j.mrfmmm.2010.10.005

103. Nencioni A, Cea M, Garuti A, Passalacqua M, Raffaghello L, Soncini D, et al. Grb7 Upregulation Is a Molecular Adaptation to HER2 Signaling Inhibition Due to Removal of Akt-Mediated Gene Repression. PLoS One. Public Library of Science; 2010;5: e9024. Available: https://doi.org/10.1371/journal.pone.0009024

104. Cancer Facts and Statistics [Internet]. 2020. Available: https://www.cancer.org/research/cancer-facts-statistics.html

105. Serag A, Ion-Margineanu A, Qureshi H, McMillan R, Saint Martin M-J, Diamond J, et al. Translational AI and Deep Learning in Diagnostic Pathology. Front Med. Frontiers Media S.A.; 2019;6: 185. doi:10.3389/fmed.2019.00185

106. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. Oxford University Press; 2018;19: 1236–1246. doi:10.1093/bib/bbx044

107. Thomsen K, Iversen L, Titlestad TL, Winther O. Systematic review of machine learning for diagnosis and prognosis in dermatology. J Dermatolog Treat. Taylor & Francis; 2019; 1–15. doi:10.1080/09546634.2019.1682500

108. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2015;521: 436. Available: https://doi.org/10.1038/nature14539

109. Kechavarzi BD, Wu H, Doman TN. Bottom-up, integrated -omics analysis identifies broadly dosage-sensitive genes in breast cancer samples from TCGA. PLoS One. Public Library of Science; 2019;14: e0210910. Available: https://doi.org/10.1371/journal.pone.0210910

110. Massey Jr. FJ. The Kolmogorov-Smirnov test for goodness of fit. J Am Stat Assoc. US: American Statistical Association; 1951;46: 68–78. doi:10.2307/2280095

111. Chollet F, others. Keras. 2015.

112. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in {P}ython. J Mach Learn Res. 2011;12: 2825–2830.

113. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg U V, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems 30. Curran Associates, Inc.; 2017. pp. 4765–4774. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

114. John JA, Draper NR. An Alternative Family of Transformations. J R Stat Soc Ser C (Applied Stat. [Wiley, Royal Statistical Society]; 1980;29: 190–197. doi:10.2307/2986305

115. Wang J, Xi X, Shang W, Acharya A, Li S, Savkovic V, et al. The molecular differences between human papillomavirus-positive and -negative oropharyngeal squamous cell carcinoma: A bioinformatics study. Am J Otolaryngol. 2019;40: 547–554. doi:https://doi.org/10.1016/j.amjoto.2019.04.015

116. Chu Y, Hu X, Wang G, Wang Z, Wang Y. Downregulation of miR-136 promotes the progression of osteosarcoma and is associated with the prognosis of patients with osteosarcoma. Oncol Lett. 2019/04/01. D.A. Spandidos; 2019;17: 5210–5218. doi:10.3892/ol.2019.10203

117. Ren H, Qi Y, Yin X, Gao J. miR-136 targets MIEN1 and involves the metastasis of colon cancer by suppressing epithelial-to-mesenchymal transition. Onco Targets Ther. Dove Medical Press; 2017;11: 67–74. doi:10.2147/OTT.S113359

118. Wang L, Chen R, Zhang Y. miR-296-3p targets APEX1 to suppress cell migration and invasion of non-small-cell lung cancer. Oncol Lett. 2019/07/05. D.A. Spandidos; 2019;18: 2612–2618. doi:10.3892/ol.2019.10572

119. Tang Y, Yang S, Wang M, Liu D, Liu Y, Zhang Y, et al. Epigenetically altered miR-193a-3p promotes HER2 positive breast cancer aggressiveness by targeting GRB7. Int J Mol Med. 2019/04/15. D.A. Spandidos; 2019;43: 2352–2360. doi:10.3892/ijmm.2019.4167

120. Costello A, Coleman O, Lao NT, Henry M, Meleady P, Barron N, et al. Depletion of endogenous miRNA-378-3p increases peak cell density of CHO DP12 cells and is correlated with elevated levels of ubiquitin carboxyl-terminal hydrolase 14. J Biotechnol. 2018;288: 30–40. doi:https://doi.org/10.1016/j.jbiotec.2018.10.008

121. Dong Z, Yu C, Rezhiya K, Gulijiahan A, Wang X. Downregulation of miR-146a promotes tumorigenesis of cervical cancer stem cells via VEGF/CDC42/PAK1 signaling pathway. Artif Cells, Nanomedicine, Biotechnol. Taylor & Francis; 2019;47: 3711–3719. doi:10.1080/21691401.2019.1664560

122. Zhou H, Cao T, Li WP, Wu G. Combined expression and prognostic significance

of PPFIA1 and ALG3 in head and neck squamous cell carcinoma. Mol Biol Rep. 2019;46: 2693–2701. doi:10.1007/s11033-019-04712-y

123. Barros-Filho MC, Reis-Rosa LA, Hatakeyama M, Marchi FA, Chulam T, Scapulatempo-Neto C, et al. Oncogenic drivers in 11q13 associated with prognosis and response to therapy in advanced oropharyngeal carcinomas. Oral Oncol. 2018;83: 81–90. doi:https://doi.org/10.1016/j.oraloncology.2018.06.010

124. Yang J, Wu N-N, Huang D-J, Luo Y-C, Huang J-Z, He H-Y, et al. PPFIA1 is upregulated in liver metastasis of breast cancer and is a potential poor prognostic indicator of metastatic relapse. Tumor Biol. SAGE Publications Ltd STM; 2017;39: 1010428317713492. doi:10.1177/1010428317713492

125. Kushwaha PP, Gupta S, Singh AK, Kumar S. Emerging Role of Migration and Invasion Enhancer 1 (MIEN1) in Cancer Progression and Metastasis [Internet]. Frontiers in Oncology . 2019. p. 868. Available: https://www.frontiersin.org/article/10.3389/fonc.2019.00868

126. Zhao H-B, Zhang X-F, Wang H-B, Zhang M-Z. Migration and invasion enhancer 1 (MIEN1) is overexpressed in breast cancer and is a potential new therapeutic molecular target. Genet Mol Res. Brazil; 2017;16. doi:10.4238/gmr16019380

127. Hu W, Wei H, Li K, Li P, Lin J, Feng R. Downregulation of USP32 inhibits cell proliferation, migration and invasion in human small cell lung cancer. Cell Prolif. John Wiley & Sons, Ltd (10.1111); 2017;50: e12343. doi:10.1111/cpr.12343

128. Zhang X, Han S, Zhou H, Cai L, Li J, Liu N, et al. TIMM50 promotes tumor progression via ERK signaling and predicts poor prognosis of non-small cell lung cancer patients. Mol Carcinog. John Wiley & Sons, Ltd; 2019;58: 767–776.

doi:10.1002/mc.22969

129. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci. 2005;102: 15545 LP – 15550. doi:10.1073/pnas.0506580102

130. Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003;34: 267–273. doi:10.1038/ng1180

131. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang T-H, et al. The Immune Landscape of Cancer. Immunity. Elsevier; 2018;48: 812-830.e14. doi:10.1016/j.immuni.2018.03.023

Curriculum Vitae

Bobak David Kechavarzi

Education

- Bachelor of Arts, Hanover College. Hanover, IN, 2007. Major: Biology, Minor: Computer Science
- Master of Science, Indiana University. Bloomington, IN, 2010. Concentration: Bioinformatics
- Doctor of Philosophy, Indiana University Purdue University Indianapolis. Indianapolis, IN, 2020. Concentration: Bioinformatics

Professional Experience

Assoc. Consultant - Informatics Capabilities

July 2013 - Ongoing

Eli Lilly and Co.,

Research Information Digital Services , Indianapolis, IN

- Developing capabilities and methodologies in Data sciences (Web crawling, interaction network analysis, heterogenous data source integration)
- Constructing storage schemas and standardization practices for -Omics data (DNA-,RNA-seq, mutation, etc) and metadata
- Engineering genomics analysis for large-scale, cross-disease comparisons
- Exploring NoSQL and high-performance databases
- Data dashboarding and searching via R Shiny and Python frameworks
- Implemented workflows for reproducible research and data discovery using document indexing and dashboarding.

- Engineered AWS-hosted collaboration spaces for external party engagement

Research Assistant

April 2012-May 2020

Indiana University Purdue University,

School of Informatics and Computing, Indianapolis, IN

- Conducted research at the Janga Lab at IUPUI (2012-2013)
- Evaluating expression patterns of RNA Binding Proteins (RBPs) in different experimental conditions
- Determining network-based methods of identifying key regulatory elements in human transcriptome.
- Ongoing research in bottom-up bioinformatics
- Machine learning for identifying sample stratification
- Using big-data, distributed computing, and systems biology concepts to facilitate hypothesis generation

Publications

- Dissecting the Expression Landscape of RNA-Binding Proteins Implicated In Human Cancers. Jan 2013 .RECOMB/ISCB Conference on Regulatory and Systems Genomics, with DREAM Challenges
- Kechavarzi, Bobak, Janga, Sarath. Dissecting the expression landscape of RNA-binding proteins in human cancers. Genome Biology 201415:R14.https://doi.org/10.1186/gb-2014-15-1-r14
- Barati MT, Powell DW, Kechavarzi BD, et al. Differential expression of endoplasmic reticulum stress-response proteins in different renal tubule subtypes

of OVE26 diabetic mice. Cell Stress Chaperones. 2016;21(1):155166. doi:10.1007/s12192-015-0648-2

- Hoffman R, Dow E, Perumal N, et al 42 Gene expression profile from 1,760 sle patients reveals novel complex interferon responsive gene networks Lupus Science Medicine 2017;4:doi: 10.1136/lupus-2017-000215.42
- Kechavarzi BD, Wu H, Doman TN (2019) Bottom-up, integrated -omics analysis identifies broadly dosage-sensitive genes in breast cancer samples from TCGA. PLOS ONE 14(1): e0210910. https://doi.org/10.1371/journal.pone.0210910